Model selection for density estimation with $\mathbb{L}_2$-loss

Lucien Birgé

Université Paris VI
Laboratoire de Probabilités et Modèles Aléatoires
U.M.R. C.N.R.S. 7599

20/10/2008

Abstract

We consider here estimation of an unknown probability density s belonging to $\mathbb{L}_2(\mu)$ where μ is a probability measure. We have at hand n i.i.d. observations with density s and use the squared $\mathbb{L}_2$-norm as our loss function. The purpose of this paper is to provide an abstract but completely general method for estimating s by model selection, allowing to handle arbitrary families of finite-dimensional (possibly non-linear) models and any $s \in \mathbb{L}_2(\mu)$. We shall, in particular, consider the cases of unbounded densities and bounded densities with unknown $\mathbb{L}_\infty$-norm and investigate how the $\mathbb{L}_\infty$-norm of s may influence the risk. We shall also provide applications to adaptive estimation and aggregation of preliminary estimators. Although of a purely theoretical nature, our method leads to results that cannot presently be reached by more concrete ones.

1 Introduction

1.1 Histograms and partition selection

Suppose we have at hand n i.i.d. observations $X_1, \ldots, X_n$ with values in the measurable space $(\mathcal{X}, \mathcal{W})$ and they have an unknown density s with respect to some probability measure μ on $\mathcal{X}$. The simplest method for finding an estimator of s is to build an histogram. Given a finite partition $\mathcal{I} = \{I_1, \ldots, I_k\}$ of $\mathcal{X}$ with $\mu(I_j) = l_j > 0$ for $1 \leq j \leq k$, the histogram $\hat{s}_{\mathcal{I}}$ based on this partition is defined by

$$\hat{s}_{\mathcal{I}}(X_1, \ldots, X_n) = \frac{1}{nl_j} \sum_{j=1}^{k} N_j \mathbb{1}_{I_j}, \quad \text{with } N_j = \sum_{i=1}^{n} \mathbb{1}_{I_j}(X_i). \tag{1.1}$$

[0] *AMS 1991 subject classifications.* Primary 62G07; secondary 62G10.

Key words and phrases. Density estimation, $\mathbb{L}_2$-loss, model selection, estimator selection, histograms.

Let

$$p_j = \int_{I_j} s\,d\mu, \quad \overline{s}_{\mathcal{I}} = \sum_{j=1}^{k} \frac{p_j}{l_j} 1\!\!1_{I_j} \quad \text{and} \quad \overline{S}_{\mathcal{I}} = \left\{ \sum_{j=1}^{k} \beta_j 1\!\!1_{I_j} \,\middle|\, \beta_j \in \mathbb{R} \text{ for } 1 \le j \le k \right\}.$$

If $s \in \mathbb{L}_2(\mu)$, then $\overline{s}_{\mathcal{I}}$ is the orthogonal projection of s onto the k-dimensional linear space $\overline{S}_{\mathcal{I}}$ spanned by the functions $1\!\!1_{I_j}$. Choosing the squared $\mathbb{L}_2$-distance induced by the norm $\|\cdot\|$ of $\mathbb{L}_2(\mu)$ as our loss function leads to the following quadratic risk for the estimator $\hat{s}_{\mathcal{I}}$:

$$\mathbb{E}\left[\|\hat{s}_{\mathcal{I}} - s\|^2\right] = \|\overline{s}_{\mathcal{I}} - s\|^2 + \frac{1}{n} \sum_{j=1}^{k} \frac{p_j(1-p_j)}{l_j}. \tag{1.2}$$

Hence, if $s \in \mathbb{L}_\infty(\mu)$, with norm $\|s\|_\infty$, the quadratic risk of $\hat{s}_{\mathcal{I}}$ can be bounded by

$$\mathbb{E}\left[\|\hat{s}_{\mathcal{I}} - s\|^2\right] \le \|\overline{s}_{\mathcal{I}} - s\|^2 + \frac{(k-1)\|s\|_\infty}{n}, \tag{1.3}$$

and, as we shall see below, this bound is essentially unimprovable without additional assumptions.

The histogram estimator $\hat{s}_{\mathcal{I}}$ is probably the simplest example of a *model-based estimator* with *model* $\overline{S}_{\mathcal{I}}$, i.e. an estimator of s with values in $\overline{S}_{\mathcal{I}}$. It may acually be viewed as the empirical counterpart of the projection $\overline{s}_{\mathcal{I}}$ of s onto $\overline{S}_{\mathcal{I}}$.

Suppose now that we are given a finite (although possibly very large) family $\{\mathcal{I}_m, m \in \mathcal{M}\}$ of finite partitions of $\mathcal{X}$ with respective cardinalities $|\mathcal{I}_m|$, hence the corresponding families of models $\{\overline{S}_{\mathcal{I}_m}, m \in \mathcal{M}\}$ and histogram estimators $\{\hat{s}_{\mathcal{I}_m}, m \in \mathcal{M}\}$. It is natural to try to find one estimator in the family which leads, at least approximately, to the minimal risk $\inf_{m \in \mathcal{M}} \mathbb{E}\left[\|\hat{s}_{\mathcal{I}_m} - s\|^2\right]$. But one cannot select such an estimator from (1.2) or (1.3) since the risk depends on the unknown density s via $\overline{s}_{\mathcal{I}_m}$. Methods of *model or estimator selection* base the choice of a suitable partition $\mathcal{I}_{\hat{m}}$ with $\hat{m} = \hat{m}(X_1, \ldots, X_n)$ on the observations. When $s \in \mathbb{L}_\infty(\mu)$ one would like to know whether it is possible to design a selection procedure $\hat{m}(X_1, \ldots, X_n)$ leading (at least approximately), in view of (1.3), to a risk bound of the form

$$\mathbb{E}\left[\|\hat{s}_{\mathcal{I}_{\hat{m}}} - s\|^2\right] \le C \inf_{m \in \mathcal{M}} \left\{ \|\overline{s}_{\mathcal{I}_m} - s\|^2 + n^{-1}\|s\|_\infty |\mathcal{I}_m| \right\},$$

for some universal constant C, even when $\|s\|_\infty$ is unknown.

1.2 What is presently known

There exists a considerable amount of litterature dealing with problems of model or estimator selection. Most of it is actually devoted to the analysis of Gaussian problems, or regression problems, or density estimation with either Hellinger or Kullback loss and it is not our aim here to review this litterature. Only a few papers are actually devoted to our subject, namely model or estimator selection for estimating densities with $\mathbb{L}_2$-loss, and we shall therefore concentrate on these only. These papers can roughly be divided into three groups: the ones dealing with penalized projection estimators, the ones that study aggregation by selection of preliminary estimators and the ones which use methods based on the thresholding of empirical coefficients within a given basis. The last ones are typically not advertised as dealing with model selection but, as explained for instance in Section 5.1.2 of Birgé and Massart (2001), they can be viewed as special instances of model selection methods for models that are spanned by some finite subsets of an orthonormal basis. All these papers have in common the fact that they require more or less severe restrictions on the families of models and, apart from some special cases, typically assume that $s \in \mathbb{L}_\infty(\mu)$ with a known or estimated bound for $\|s\|_\infty$.

In order to see how such methods apply to our problem of partition selection, let us be more specific and assume that $\mathscr{X} = [0,1]$, μ is the Lebesgue measure and $\mathscr{N} = \{j/(N+1), 1 \leq j \leq N\}$ for some (possibly very large) positive integer N. For any subset m of $\mathscr{N}$, we denote by $\mathcal{I}_m$ the partition of $\mathscr{X}$ generated by the intervals with set of endpoints $m \cup \{0,1\}$ and we set $\overline{S}_m = \overline{S}_{\mathcal{I}_m}$ and $\hat{s}_m = \hat{s}_{\mathcal{I}_m}$. This leads to a set $\mathcal{M}$ with cardinality 2^N and the corresponding families of linear models $\left\{\overline{S}_m, m \in \mathcal{M}\right\}$ and related histogram estimators $\{\hat{s}_m, m \in \mathcal{M}\}$. Then all models $\overline{S}_m$ are linear subspaces of the largest one $\overline{S}_{\mathscr{N}}$. Of particular interest is the dyadic case with $N = 2^K - 1$ for which $\overline{S}_{\mathscr{N}}$ is the linear span of the 2^K first coefficients of the Haar basis. There is, nevertheless, a difference between expansions in the Haar basis and projections on our family of models. Let us, for instance, consider the function $1\!\!1_{[0,2^{-K})}$. It belongs to the two-dimensional model $\overline{S}_{\{1\}}$ but its expansion in the Haar basis has K non-zero coefficients.

Given a sample $X_1, \ldots, X_n$ with unknown density s, which partition $\mathcal{I}_m$ should we choose to estimate s and what bound could we derive for the resulting estimator? Penalized projection estimators have been considered by Birgé and Massart (1997) and an improved version is to be found in Chapter 7 of Massart (2007). The method either deals with polynomial collections of models (which does not apply to our case) or with subset selection within a given basis which applies here only when $N = 2^K - 1$ and we use the Haar basis. Moreover, it requires that $N < n/\log n$ and a bound on $\|\overline{s}_{\mathcal{I}_{\mathscr{N}}}\|_\infty$ be known or estimated, as in Section 4.4.4 of Birgé and Massart (1997), since the penalty depends on it.

Methods based on wavelet thresholding, as described in Donoho, Johnstone, Kerkyacharian and Picard (1996) or Kerkyacharian and Picard (2000) (see also the numerous references therein) require the same type of restrictions and, in particular, a control on $\|s\|_\infty$ in order to properly calibrate the threshold. Also, as mentioned above, restricting to subsets of the Haar basis may result in expansions that use many more coefficients (K instead of 2, for instance) than needed with the partition selection approach.

Aggregation of estimators by selection assumes that preliminary estimators (one for each model in our case) are given in advance (we should here use the histograms) and typically leads to a risk bound including a term of the form $n^{-1}\|s\|_\infty \log|\mathcal{M}| = n^{-1}N\|s\|_\infty \log 2$ so that all such results are useless for $N \geq n$. Moreover, most of them also require that an upper bound for $\|s\|_\infty$ be known since it enters the construction of the aggregate estimator. This is the case in Rigollet (2006) (see for instance his Corollary 2.7) and Juditsky, Rigollet and Tsybakov (2007, Corollary 5.7) since the parameter β that governs their mirror averaging method depends crucially on an upper bound for $\|s\|_\infty$. As to Samarov and Tsybakov (2005), their Assumption 1 requires that N be not larger than $C\log n$. Similar restrictions are to be found in

Yang (2000) in his developments for mixing strategies and in Rigollet and Tsybakov (2007) for linear aggregation of estimators. Lounici (2008) does not assume that $s \in \mathbb{L}_\infty$ but, instead, that all preliminary estimators are uniformly bounded. One can always truncate the estimators to get this, but to be efficient, the truncation should be adapted to the unknown parameter s, and therefore chosen from the data in a suitable way. We do not know of any paper that allows such a data driven choice.

Consequently, none of these results can solve our partition selection problem in a satisfactory way when N is at least of size n and $\|s\|_\infty$ is unknown. This fact was one motivation for our study of model selection for density estimation with $\mathbb{L}_2$-loss. Results about partition selection will be a consequence of our general treatment of model selection. This treatment allows to consider arbitrary countable families of finite-dimensional models (possibly nonlinear) and does not put any assumption on the density s, apart from the fact that it belongs to $\mathbb{L}_2(\mu)$; it may, in particular, be unbounded. We do not know of any result that applies to such a situation. There is a counterpart to this level of generality: our procedure is of a purely abstract nature and not constructive, only indicating what is theoretically feasible. Unfortunately, we are unable to design a practical procedure with similar properties.

2 Model based estimation and model selection

To begin with, let us fix our framework and notations. We want to estimate an unknown density s, with respect to some *probability* measure μ on the measurable space $(\mathscr{X}, \mathscr{W})$, from an i.i.d. sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ of random variables $X_i \in \mathscr{X}$ with distribution $P_s = s \cdot \mu$. Throughout the paper we denote by $\mathbb{P}_s$ the probability that gives $\boldsymbol{X}$ the distribution $P_s^{\otimes n}$, by $\mathbb{E}_s$ the corresponding expectation operator and by $\|\cdot\|_q$ the norm in $\mathbb{L}_q(\mu)$, omitting the subscript when $q = 2$ for simplicity. We denote by d_2 the distance in $\mathbb{L}_2(\mu)$: $d_2(t, u) = \|t - u\|$. For $1 \le q \le +\infty$ and $\Gamma > 1$, we set

$$\overline{\mathbb{L}}_q = \left\{ t \in \mathbb{L}_q(\mu) \;\middle|\; t \ge 0 \text{ and } \int t\, d\mu = 1 \right\}; \quad \overline{\mathbb{L}}_\infty^\Gamma = \left\{ t \in \overline{\mathbb{L}}_\infty \mid \|t\|_\infty \le \Gamma \right\}. \quad (2.1)$$

We measure the performance at $s \in \overline{\mathbb{L}}_2$ of an estimator $\hat{s}(\boldsymbol{X}) \in \mathbb{L}_2$ by its quadratic risk $\mathbb{E}_s\left[d_2^2\left(\hat{s}(\boldsymbol{X}), s\right)\right]$. More generally, if (M, d) is a metric space of measurable functions on $\mathscr{X}$ such that $M \cap \overline{\mathbb{L}}_1 \neq \emptyset$, the quadratic risk of some estimator $\hat{s} \in M$ at $s \in M \cap \overline{\mathbb{L}}_1$ is defined as $\mathbb{E}_s\left[d^2\left(\hat{s}(\boldsymbol{X}), s\right)\right]$. We denote by $|\mathcal{I}|$ the cardinality of the set $\mathcal{I}$ and set $a \vee b$ and $a \wedge b$ for the maximum and the minimum of a and b, respectively. Throughout the paper C (or C', ...) will denote a universal (numerical) constant and $C(a, b, \ldots)$ or C_q a fonction of the parameters $a, b, \ldots$ or q. Both may vary from line to line. Finally, from now on, *countable* will always mean "finite or countable".

2.1 Model based estimation

A common method for estimating s consists in choosing a particular subset $\overline{S}$ of (M, d) that we shall call a *model* for s and design an estimator with values in $\overline{S}$. Of this type are the *maximum likelihood estimator* over $\overline{S}$ or the *projection estimator* onto $\overline{S}$. Let us set $M = \overline{\mathbb{L}}_1$ and choose for d either the Hellinger distance h or the variation distance v given respectively by

$$h^2(t,u) = \frac{1}{2}\int\left(\sqrt{t}-\sqrt{u}\right)^2 d\mu \qquad \text{and} \qquad v(t,u) = \frac{1}{2}\int|t-u|\,d\mu.$$

It follows from Le Cam (1973, 1975, 1986) and subsequent results by Birgé (1983, 2006a) that the risk of suitably designed estimators with values in $\overline{S}$ is the sum of two terms, an *approximation term* depending on the distance from s to $\overline{S}$ and an *estimation term* depending on the dimension of the model $\overline{S}$ which can be defined as follows.

Definition 1 *Let $\overline{S}$ be a subset of some metric space (M,d) and let $\mathcal{B}_d(t,r)$ denote the open ball of center t and radius r with respect to the metric d. Given $\eta > 0$, a subset S_η of M is called an η-net for $\overline{S}$ if, for each $t \in \overline{S}$, one can find $t' \in S_\eta$ with $d(t,t') \leq \eta$.*

We say that $\overline{S}$ has a metric dimension bounded by $D \geq 0$ *if, for every $\eta > 0$, there exists an η-net S_η for $\overline{S}$ such that*

$$|S_\eta \cap \mathcal{B}_d(t, x\eta)| \leq \exp\left[Dx^2\right] \quad \textit{for all } x \geq 2 \textit{ and } t \in M. \tag{2.2}$$

Remark: *One can always assume that $S_\eta \subset \overline{S}$ at the price of replacing D by $25D/4$ according to Proposition 7 of Birgé (2006a).*

Typical examples of sets with metric dimension bounded by D when (M,d) is a normed linear space are subsets of $2D$-dimensional linear subspaces of $\mathcal{M}$ as shown in Birgé (2006a). If d is either h or v and $\overline{S} \subset \overline{\mathbb{L}}_1$ has a metric dimension bounded by $D \geq 1/2$, there exists a universal constant C and an estimator $\hat{s}(\boldsymbol{X})$ with values in $\overline{S}$ such that, for any $s \in \overline{\mathbb{L}}_1$,

$$\mathbb{E}_s\left[d^2\left(\hat{s}(\boldsymbol{X}), s\right)\right] \leq C\left[\inf_{t\in\overline{S}} d^2(s,t) + n^{-1}D\right]. \tag{2.3}$$

In particular, $\sup_{s\in\overline{S}}\mathbb{E}_s\left[d^2\left(\hat{s}(\boldsymbol{X}),s\right)\right] \leq Cn^{-1}D$. This results from the following theorem about model selection of Birgé (2006a) by setting $\mathcal{M} = \{0\}$, $\overline{S}_0 = \overline{S}$, $D_0 = D$ and $\Delta_0 = 1/2$.

Theorem 1 *Let $X_1,\ldots,X_n$ be an i.i.d. sample with unknown density s belonging to $\overline{\mathbb{L}}_1$ and $\left\{\overline{S}_m, m \in \mathcal{M}\right\}$ a finite or countable family of subsets of $\overline{\mathbb{L}}_1$ with metric dimensions bounded by $D_m \geq 1/2$ respectively. Let the nonegative weights $\Delta_m, m \in \mathcal{M}$ satisfy*

$$\sum_{m\in\mathcal{M}} \exp[-\Delta_m] = \Sigma < +\infty. \tag{2.4}$$

Then there exists a universal constant C and an estimator $\tilde{s}(X_1,\ldots,X_n)$ such that, for any $s \in \overline{\mathbb{L}}_1$,

$$\mathbb{E}_s\left[d^2\left(\tilde{s},s\right)\right] \leq C(1+\Sigma)\inf_{m\in\mathcal{M}}\left[\inf_{t\in\overline{S}_m} d^2(s,t) + n^{-1}(D_m \vee \Delta_m)\right]. \tag{2.5}$$

Unfortunately, as we shall see below, (2.3) does not hold in general when $(M,d) = (\mathbb{L}_2(\mu), d_2)$. In particular, whatever the estimator $\hat{s}$, $\sup_{s\in\overline{S}} \mathbb{E}_s\left[d^2\left(\hat{s}(\boldsymbol{X}), s\right)\right]$ may be infinite even if $\overline{S} \subset \mathbb{L}_2(\mu)$ has a bounded metric dimension. This difference is due to the following fact: h and v are actually distances defined on the set of all probabilities on $(\mathcal{X}, \mathcal{W})$ and $h(s,t) = h(P_s, P_t)$ is independent of the choice of the underlying dominating measure, the same property holding for the variation distance v. This is not the case for d_2 which is a distance in $\mathbb{L}_2(\mu)$ depending on the choice of μ and definitely not a distance between probabilities. Even the fact that $s = dP_s/d\mu$ belong or not to $\mathbb{L}_2(\mu)$ depends on μ. Further remarks on this subject can be found in Devroye and Györfi (1985) and Devroye (1987).

Nevertheless, the $\mathbb{L}_2$-distance has been much more popular in the past than either the Hellinger or variation distances, mainly because of its simplicity due to the classical "squared bias plus variance" decomposition of the risk. Although hundreds of papers have been devoted to the derivation of risk bounds for various specific estimators, we do not know of any general bound for the risk similar to (2.3) based on purely metric considerations for the distance d_2.

2.2 Projection and histogram estimators

To illustrate the specificity of the $\mathbb{L}_2$-risk, let us turn to a quite classical family of model-based estimators for densities, the projection estimators of Cencov (1962). To estimate a density $s \in \overline{\mathbb{L}}_2$ from an i.i.d. sample $X_1, \ldots, X_n$, we chose some k-dimensional linear subspace $\overline{S}$ of $\mathbb{L}_2(\mu)$ together with an orthonormal basis $(\varphi_1, \ldots, \varphi_k)$ so that the projection $\overline{s}$ of s onto $\overline{S}$ can be written $\overline{s} = \sum_{j=1}^k \beta_j\varphi_j$. Then we estimate each coefficient $\beta_j = \int \varphi_j s\, d\mu$ in this expansion by its empirical version $\hat{\beta}_j = n^{-1}\sum_{i=1}^n \varphi_j(X_i)$. This results in the projection estimator $\hat{s} = \sum_{j=1}^k \hat{\beta}_j\varphi_j$ (which in general does not belong to $\overline{\mathbb{L}}_1$) with risk

$$\begin{aligned}
\mathbb{E}_s\left[\|\hat{s} - s\|^2\right] &= \|\overline{s} - s\|^2 + n^{-1}\sum_{j=1}^k \mathrm{Var}_s\left(\varphi_j(X_1)\right) \\
&\leq \|\overline{s} - s\|^2 + n^{-1}\int\left[\sum_{j=1}^k \varphi_j^2(x)\right] s(x)\, d\mu(x) \\
&\leq \|\overline{s} - s\|^2 + n^{-1}\min\left\{\left\|\sum_{j=1}^k \varphi_j^2\right\|_\infty ; k\|s\|_\infty\right\}. \qquad (2.6)
\end{aligned}$$

A particular case occurs with the histogram $\hat{s}_{\mathcal{I}}$ given by (1.1) which corresponds to choosing $\varphi_j = l_j^{-1/2}\mathbb{1}_{I_j}$, $\overline{S} = \overline{S}_{\mathcal{I}}$ and $\overline{s} = \overline{s}_{\mathcal{I}}$. If $l_j = k^{-1}$ for all j, we get a *regular* histogram and derive from (1.2) and a convexity argument that

$$\mathbb{E}_s\left[\|s - \hat{s}_{\mathcal{I}}\|^2\right] \leq \|s - \overline{s}_{\mathcal{I}}\|^2 + (k-1)/n.$$

But, for general partitions, the bound (1.3) clearly emphasizes the difference with the risk bound of the form (2.3) obtained in Birgé and Rozenholc (2006) for the Hellinger loss:

$$\mathbb{E}_s\left[h^2(s, \hat{s}_{\mathcal{I}})\right] \leq h^2(s, \overline{s}_{\mathcal{I}}) + (k-1)/(2n).$$

Moreover, (1.3) is essentially unimprovable without further assumptions on s if the partition $\mathcal{I}$ is arbitrary, as shown by the following example. Define the partition $\mathcal{I}$ on $\mathcal{X} = [0,1]$ by $I_j = [(j-1)\alpha, j\alpha)$ for $1 \leq j < k$ and $I_k = [(k-1)\alpha, 1]$ with $0 < \alpha < (k-1)^{-1}$. Set $s = s_{\mathcal{I}} = [(k-1)\alpha]^{-1}\left[1 - 1\!\!1_{I_k}\right]$. Then $p_j = (k-1)^{-1}$ for $1 \leq j < k$, $s = \overline{s}_{\mathcal{I}}$ and it follows from (1.2) that

$$\mathbb{E}_s\left[\|s - \hat{s}_{\mathcal{I}}\|^2\right] = \frac{k-2}{(k-1)\alpha n} = \frac{(k-2)\|s\|_\infty}{n},$$

which shows that there is little space for improvement in (1.3).

2.3 Some negative results

The fact that the $\mathbb{L}_\infty$-norm of s comes into the risk is not due to the use of histograms or projection estimators as shown by another negative result provided by Proposition 4 of Birgé (2006b) that we recall below for the sake of completeness.

Proposition 1 *For each $L > 0$ and each integer D with $1 \leq D \leq 3n$, one can find a finite set $\overline{S}$ of densities with the following properties:*

i) it is a subset of some D-dimensional affine subspace of $\mathbb{L}_2([0,1], dx)$ with a metric dimension bounded by $D/2$;

ii) $\sup_{s\in\overline{S}} \|s\|_\infty \leq L + 1$;

iii) for any estimator $\hat{s}(X_1, \ldots, X_n)$ belonging to $\mathbb{L}_2([0,1], dx)$ and based on an i.i.d. sample with density $s \in \overline{S}$,

$$\sup_{s\in\overline{S}} \mathbb{E}_s\left[\|\hat{s} - s\|^2\right] > 0.0139DLn^{-1}. \tag{2.7}$$

It follows that there is no hope to get an analogue of (2.3), under the same assumptions, when $d = d_2$ and the best one can expect in full generality, when $\overline{S}$ is a model with metric dimension bounded by D and $s \in \mathbb{L}_\infty$, is to design an estimator $\hat{s}$ with a risk bounded by

$$\mathbb{E}_s\left[d_2^2(\hat{s}, s)\right] \leq C\left[\inf_{t\in\overline{S}} d_2^2(s,t) + n^{-1}D\|s\|_\infty\right]. \tag{2.8}$$

The situation becomes worse when $s \notin \mathbb{L}_\infty(\mu)$ or if $\sup_{s\in\overline{S}} \|s\|_\infty = +\infty$ as shown by the following lower bound to be proved in Section 7.1.

Proposition 2 *Let $\overline{S} = \{s_\theta, 0 < \theta \leq 1/3\}$ be the set of densities with respect to Lebesgue measure on $[0,1]$ given by*

$$s_\theta = \theta^{-2} 1\!\!1_{[0,\theta^3]} + \left(\theta^2 + \theta + 1\right)^{-1} 1\!\!1_{(\theta^3,1]}.$$

If we have at disposal n i.i.d. observations with density $s_\theta \in \overline{S}$, we can build an estimator $\hat{s}_n$ such that $\sup_{0<\theta\leq 1/3} \mathbb{E}_{s_\theta}\left[nh^2(s_\theta, \hat{s}_n)\right] \leq C$ for some C independent of n. On the other hand, although the metric dimension of $\overline{S}$ with respect to the distance d_2 is bounded by 2, $\sup_{0<\theta\leq 1/3} \mathbb{E}_{s_\theta}\left[\|s_\theta - \tilde{s}_n\|^2\right] = +\infty$, whatever n and the estimator $\tilde{s}_n$.

2.4 About this paper

Our purpose here is twofold. We first want to derive estimators achieving a risk bound which generalizes (2.8) in the sense that it could also apply to the case of $s \not\in \mathbb{L}_\infty$. We know from (2.6) that projection estimators do satisfy (2.8) when $\overline{S}$ is a D-dimensional linear space, but do not have any result for non-linear models. Our second goal is to handle many models simultaneously and design an estimator which performs as well (or almost as well) as the estimator based on the "best model", i.e. one leading to the smallest risk bound (up to some universal constant). This is possible when the distance d is either the Hellinger distance h or the variation distance v, as shown by Theorem 1. As compared to the bound (2.3), we simply pay the price of replacing D_m by Δ_m when $\Delta_m > D_m$. This price is due to the complexity of the family of models we use (there is nothing to pay in the simplest case of one model per dimension) and this price is essentially unavoidable, as shown in a specific case by Birgé and Massart (2006).

It follows from the previous section that it is impossible to get an analogue of Theorem 1 when $d = d_2$. We shall explain what kind of (necessarily weaker) results can be obtained in this context and to what extent Theorem 1 can be rescued. For this, we shall proceed in several steps. In the next section we shall explain how to build estimators based on families of special models S_m, following the method explained in Birgé (2006a). These models need to be discrete subsets of $\overline{\mathbb{L}}_\infty^\Gamma$ (for some given Γ) with bounded metric dimension while there is no reason that our initial models $\overline{S}_m$ be of this type (think of linear models). Section 4 will therefore be devoted to the construction of such special models S_m from ordinary ones. This construction will finally lead to an estimator $\hat{s}^\Gamma$ belonging to $\overline{\mathbb{L}}_\infty^\Gamma$, the performance of which strongly depends on our choice of Γ. In Section 5, we shall explain how, given a geometrically increasing sequence $(\Gamma_i)_{i\geq 1}$ of values of Γ and the corresponding sequence of estimators $\hat{s}^{\Gamma_i}$, we can use the observations to choose a suitable value for Γ. Since we have a single sample $\boldsymbol{X}$ to build the estimators $\hat{s}^{\Gamma_i}$ and to choose Γ, we shall proceed by sample splitting using one half of the sample for the construction of the estimators and the second half to select a value of Γ. In particular, for the case of a single model, this will lead to a generalized version of (2.8) that can also handle the case of $s \not\in \mathbb{L}_\infty$. When $s \in \mathbb{L}_\infty$ (with an unknown value of $\|s\|_\infty$), the risk bounds we get completely parallel (apart from some constants depending on $\|s\|_\infty$) those obtained for estimating s in the white noise model. We shall give a few applications of these results, in particular to aggregation of preliminary estimators, in Section 6, while the last section will be devoted to the most technical proofs.

3 T-estimators for $\mathbb{L}_2$-loss

In order to define estimators based on families of models with bounded metric dimensions, we shall follow the approach of Birgé (2006a) based on what we have called T-estimators. We refer to this paper for the definition of these estimators, recalling that it relies on the existence of suitable tests between balls of the underlying metric space $\left(\overline{\mathbb{L}}_2, d_2\right)$. To derive such tests, we need a few specific technical tools to deal with the $\mathbb{L}_2$-distance.

3.1 Tests between $\mathbb{L}_2$-balls

3.1.1 Randomizing our sample

In the sequel we shall make use of randomized tests based on a randomization trick due to Yang and Barron (1998, page 106) which has the effect of replacing all densities involved in our problem by new ones which are uniformly bounded away from zero. For this, we choose some number $\lambda \in (0,1)$ and consider the mapping τ from $\overline{\mathbb{L}}_2$ to $\overline{\mathbb{L}}_2$ given by $\tau(u) = \lambda u + 1 - \lambda$. Note that τ is one-to-one and isometric, up to a factor λ, i.e. $d_2(\tau(u), \tau(v)) = \lambda d_2(u,v)$. If $u \in \overline{\mathbb{L}}_\infty^\Gamma$, then $\tau(u) \in \overline{\mathbb{L}}_\infty^{\Gamma'}$ with $\Gamma' = \lambda\Gamma + 1 - \lambda$.

Let $s' = \tau(s)$. Given our initial i.i.d. sample $\boldsymbol{X}$, we want to build new i.i.d. variables $X'_1, \ldots, X'_n$ with density s'. For this, we consider two independent n-samples, $Z_1, \ldots, Z_n$ and $\varepsilon_1, \ldots, \varepsilon_n$ with respective distributions μ and Bernoulli with parameter λ. Both samples are independent of $\boldsymbol{X}$. We then set $X'_i = \varepsilon_i X_i + (1 - \varepsilon_i) Z_i$ for $1 \leq i \leq n$. It follows that X'_i has density s' as required. We shall still denote by $\mathbb{P}_s$ the probability on Ω that gives $\boldsymbol{X}' = (X'_1, \ldots, X'_n)$ the distribution $P_{s'}^{\otimes n}$. Given two distinct points $t, u \in \overline{\mathbb{L}}_2$ we define a (randomized) test function $\psi(\boldsymbol{X}')$ between t and u as a measurable function with values in $\{t, u\}$, $\psi(\boldsymbol{X}') = t$ meaning deciding t and $\psi(\boldsymbol{X}') = u$ meaning deciding u.

Once we have used the randomization trick of Yang and Barron, for instance with $\lambda = 1/2$, we deal with an i.i.d. sample $\boldsymbol{X}'$ with a density s' which is bounded from below by $1/2$ and we may therefore work within the set of densities that satisfy this property.

3.1.2 Some minimax results

The main tool for the design of tests between $\mathbb{L}_2$-balls of densities is the following proposition which derives from the results of Birgé (1984) (keeping here the notations of that paper) and in particular from Corollary 3.2, specialized to the case of $I = \{t\}$ and $c = 0$.

Proposition 3 *Let $\mathcal{M}$ be some linear space of finite measures on some measurable space $(\Omega, \mathcal{A})$ with a topology of a locally convex separated linear space. Let $\mathcal{P}, \mathcal{Q}$ be two disjoint sets of probabilities in $\mathcal{M}$ and F a set of positive measurable functions on Ω with the following properties (with respect to the given topology on $\mathcal{M}$):*

i) $\mathcal{P}$ and $\mathcal{Q}$ are convex and compact;

ii) for any $f \in F$ and $0 < z < 1$ the function $P \mapsto \int f^z \, dP$ is well-defined and upper semi-continuous on $\mathcal{P} \cup \mathcal{Q}$;

iii) for any $P \in \mathcal{P}$, $Q \in \mathcal{Q}$, $t \in (0,1)$ and $\varepsilon > 0$, there exists an $f \in F$ such that

$$(1-t) \int f^t \, dP + t \int f^{1-t} \, dQ < \int (dP)^{1-t} (dQ)^t + \varepsilon;$$

iv) all probabilities in $\mathcal{P}$ (respectively in $\mathcal{Q}$) are mutually absolutely continuous. Then one can find $\overline{P} \in \mathcal{P}$ and $\overline{Q} \in \mathcal{Q}$ such that

$$\begin{aligned}\sup_{P \in \mathcal{P}} \int \left(\frac{\overline{Q}}{\overline{P}}\right)^t dP \; &= \; \sup_{Q \in \mathcal{Q}} \int \left(\frac{\overline{P}}{\overline{Q}}\right)^{1-t} dQ \; = \; \sup_{P \in \mathcal{P}, Q \in \mathcal{Q}} \int (dP)^{1-t} (dQ)^t \\ &= \int \left(d\overline{P}\right)^{1-t} \left(d\overline{Q}\right)^t .\end{aligned}$$

In Birgé (1984) we assumed that $\mathcal{M}$ was the set of *all* finite measures on $(\Omega, \mathcal{A})$ but the proof actually only uses the fact that $\mathcal{P}$ and $\mathcal{Q}$ are subsets of $\mathcal{M}$. Recalling that the Hellinger affinity between two densities u and v is defined by $\rho(u,v) = \int \sqrt{uv}\, d\mu = 1 - h^2(u,v)$, we get the following corollary.

Corollary 1 *Let μ be a probability measure on $(\mathcal{X}, \mathcal{W})$ and, for $1 \leq i \leq n$, let $(\mathcal{P}_i, \mathcal{Q}_i)$ be a pair of disjoint convex and weakly compact subsets of $\mathbb{L}_2(\mu)$ such that*

$$s > 0 \;\; \mu\text{-a.s.} \qquad \text{and} \qquad \int s\, d\mu = 1 \quad \text{for all } s \in \bigcup_{i=1}^{n} (\mathcal{P}_i \cup \mathcal{Q}_i)\,. \tag{3.1}$$

For each i, one can find $p_i \in \mathcal{P}_i$ and $q_i \in \mathcal{Q}_i$ such that

$$\sup_{u \in \mathcal{P}_i} \int \sqrt{q_i/p_i}\, u\, d\mu = \sup_{v \in \mathcal{Q}_i} \int \sqrt{p_i/q_i}\, v\, d\mu = \sup_{u \in \mathcal{P}_i, v \in \mathcal{Q}_i} \rho(u,v) = \rho(p_i, q_i).$$

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random vector on $\mathcal{X}^n$ with distribution $\bigotimes_{i=1}^n (s_i \cdot \mu)$ with $s_i \in \mathcal{P}_i$ for $1 \leq i \leq n$ and let $x \in \mathbb{R}$. Then

$$\mathbb{P}\left[\sum_{i=1}^n \log(q_i/p_i)(X_i) \geq 2x\right] \leq e^{-x} \prod_{i=1}^n \rho(p_i, q_i) \leq \exp\left[-x - \sum_{i=1}^n h^2(p_i, q_i)\right].$$

If $\boldsymbol{X}$ has distribution $\bigotimes_{i=1}^n (u_i \cdot \mu)$ with $u_i \in \mathcal{Q}_i$ for $1 \leq i \leq n$, then

$$\mathbb{P}\left[\sum_{i=1}^n \log(q_i/p_i)(X_i) \leq 2x\right] \leq e^{x} \prod_{i=1}^n \rho(p_i, q_i) \leq \exp\left[x - \sum_{i=1}^n h^2(p_i, q_i)\right].$$

Proof: We apply the previous proposition with $t = 1/2$, $(\mathcal{X}, \mathcal{W}) = (\Omega, \mathcal{A})$ and $\mathcal{M}$ the set of measures of the form $u \cdot \mu$, $u \in \mathbb{L}_2(\mu)$ endowed with the weak $\mathbb{L}_2$-topology. In view of (3.1), $\mathcal{P}_i$ and $\mathcal{Q}_i$ can be identified with two sets of probabilities and we can take for F the set of all positive functions such that $\log f$ is bounded. As a consequence, all four assumptions of Proposition 3 are satisfied. In order to get iii) we simply take for f a suitably truncated version of s/u when $P = s \cdot \mu$ and $Q = u \cdot \mu$. As to the probability bounds they derive from classical exponential inequalities, as for Lemma 7 of Birgé (2006a). □

3.1.3 Abstract tests between $\mathbb{L}_2$-balls

The purpose of this section is to prove the following result.

Theorem 2 *Let $t, u \in \overline{\mathbb{L}}_\infty^\Gamma$ for some $\Gamma < +\infty$. For any $x \in \mathbb{R}$, there exists a test $\psi_{t,u,x}$ between t and u, based on the randomized sample $\boldsymbol{X}'$ defined in Section 3.1.1 with a suitable value of λ, which satisfies*

$$\sup_{\left\{s \in \overline{\mathbb{L}}_2 \,|\, d_2(s,t) \leq d_2(t,u)/4\right\}} \mathbb{P}_s[\psi_{t,u,x}(\boldsymbol{X}') = u] \leq \exp\left[-\frac{n\left(\|t-u\|^2 + x\right)}{65\Gamma}\right], \tag{3.2}$$

$$\sup_{\left\{s \in \overline{\mathbb{L}}_2 \,|\, d_2(s,u) \leq d_2(t,u)/4\right\}} \mathbb{P}_s[\psi_{t,u,x}(\boldsymbol{X}') = t] \leq \exp\left[-\frac{n\left(\|t-u\|^2 - x\right)}{65\Gamma}\right]. \tag{3.3}$$

Proof: It requires several steps. To begin with, we use the randomization trick of Yang and Barron described in Section 3.1.1, replacing our original sample $\boldsymbol{X}$ by the randomized sample $\boldsymbol{X}' = (X'_1, \ldots, X'_n)$ for some convenient value of λ to be chosen later. Each X'_i has density $s' \geq 1 - \lambda$ when X_i has density s. Then we build a test between $t' = \tau(t)$ and $u' = \tau(u)$ based on $\boldsymbol{X}'$ and Corollary 1. To do this, we set $\Delta = \|t - u\|$,

$$\mathcal{P} = \tau\left(\mathcal{B}_{d_2}(t, \Delta/4) \cap \overline{\mathbb{L}}_2\right) \qquad \text{and} \qquad \mathcal{Q} = \tau\left(\mathcal{B}_{d_2}(u, \Delta/4) \cap \overline{\mathbb{L}}_2\right).$$

Then $\mathcal{P}$ is the subset of the ball $\mathcal{B}_{d_2}(t', \lambda\Delta/4)$ of those densities bounded from below by $1 - \lambda$, hence $d_2(\mathcal{P}, \mathcal{Q}) \geq \lambda\Delta/2$ and $\mathcal{P}$ is convex and weakly closed since any indicator function belongs to $\mathbb{L}_2(\mu)$ because μ is a probability. Since $\mathcal{B}_{d_2}(t', \lambda\Delta/4)$ is weakly compact, it is also the case for $\mathcal{P}$ and the same argument shows that $\mathcal{Q}$ is also convex and weakly compact. It then follows from Corollary 1 that one can find $\bar{t} \in \mathcal{P}$ and $\bar{u} \in \mathcal{Q}$ such that

$$\mathbb{P}_s\left[\sum_{i=1}^{n} \log\left(\bar{u}(X'_i)/\bar{t}(X'_i)\right) \geq 2y\right] \leq \exp\left[-nh^2\left(\bar{t}, \bar{u}\right) - y\right] \quad \text{if } s \in \mathcal{P}, \tag{3.4}$$

while

$$\mathbb{P}_s\left[\sum_{i=1}^{n} \log\left(\bar{u}(X'_i)/\bar{t}(X'_i)\right) \leq 2y\right] \leq \exp\left[-nh^2\left(\bar{t}, \bar{u}\right) + y\right] \quad \text{if } s \in \mathcal{Q}. \tag{3.5}$$

Fixing $y = nx/(65\Gamma)$, we finally define $\psi_{t,u,x}(\boldsymbol{X}')$ by setting $\psi_{t,u,x}(\boldsymbol{X}') = u$ if and only if $\sum_{i=1}^{n} \log\left(\bar{u}(X'_i)/\bar{t}(X'_i)\right) \geq 2y$. Since $s' \in \mathcal{P}$ is equivalent to $s \in \mathcal{B}_{d_2}(t, \Delta/4)$ or $d_2(s,t) \leq \Delta/4$ and similarly $s \in \mathcal{Q}$ is equivalent to $d_2(s,u) \leq \Delta/4$, to derive (3.2) and (3.3) from (3.4) and (3.5), we just have to show that $h^2\left(\bar{t}, \bar{u}\right) \geq (65\Gamma)^{-1}\Delta^2$. We start from the fact, to be proved below, that

$$\|\bar{t} \vee \bar{u}\|_\infty \leq 2(\lambda\Gamma + 1 - \lambda). \tag{3.6}$$

It implies that

$$\begin{aligned} h^2\left(\bar{t}, \bar{u}\right) &= \frac{1}{2}\int\left(\sqrt{\bar{t}} - \sqrt{\bar{u}}\right)^2 d\mu = \frac{1}{2}\int \frac{\left(\bar{t} - \bar{u}\right)^2}{\left(\sqrt{\bar{t}} + \sqrt{\bar{u}}\right)^2} d\mu \\ &\geq \frac{\|\bar{t} - \bar{u}\|^2}{16(\lambda\Gamma + 1 - \lambda)} \geq \frac{(\lambda\Delta)^2}{64(\lambda\Gamma + 1 - \lambda)}. \end{aligned}$$

Choosing λ close enough to one leads to the required bound $h^2\left(\bar{t}, \bar{u}\right) \geq (65\Gamma)^{-1}\Delta^2$. As to (3.6), it is a consequence of the next lemma to be proved in Section 7.2. We apply this lemma to the pair t', u' which satisfies $\|t' \vee u'\|_\infty \leq \lambda\Gamma + 1 - \lambda$. If (3.6) were wrong, we could find $\bar{t}' \in \mathcal{P}$ and $\bar{u}' \in \mathcal{Q}$ with $h\left(\bar{t}', \bar{u}'\right) < h\left(\bar{t}, \bar{u}\right)$, which, by Corollary 1, is impossible. □

Lemma 1 *Let us consider four elements t, u, v_1, v_2 in $\overline{\mathbb{L}}_2$ with $t \neq u$, $v_1 \neq v_2$ and $\|t \vee u\|_\infty = B$. If $\|v_1 \vee v_2\|_\infty > 2B$, there exists $v'_1, v'_2 \in \overline{\mathbb{L}}_2$ with $d_2(v'_1, t) \leq d_2(v_1, t)$, $d_2(v'_2, u) \leq d_2(v_2, u)$ and $h(v'_1, v'_2) < h(v_1, v_2)$.*

3.2 The performance of T-estimators

We are now in a position to prove an analogue of Corollary 6 of Birgé (2006a).

Theorem 3 *Assume that we observe n i.i.d. random variables with unknown density $s \in \left(\overline{\mathbb{L}}_2, d_2\right)$ and that we have at disposal a countable family of discrete subsets $\{S_m\}_{m\in\mathcal{M}}$ of $\overline{\mathbb{L}}_\infty^\Gamma$ for some given $\Gamma > 1$. Let each set S_m satisfy*

$$|S_m \cap \mathcal{B}_{d_2}(t, x\eta_m)| \leq \exp\left[D_m x^2\right] \quad \textit{for all } x \geq 2 \textit{ and } t \in \overline{\mathbb{L}}_2, \tag{3.7}$$

with $\eta_m > 0$, $D_m \geq 1/2$,

$$\eta_m^2 \geq \frac{273\Gamma D_m}{n} \quad \textit{for all } m \in \mathcal{M}, \qquad \sum_{m\in\mathcal{M}} \exp\left[-\frac{n\eta_m^2}{1365\Gamma}\right] = \Sigma < +\infty. \tag{3.8}$$

Then one can build a T-estimator $\hat{s}$ such that, for all $s \in \overline{\mathbb{L}}_2$,

$$\mathbb{E}_s\left[d_2^q(s,\hat{s})\right] \leq C_q(\Sigma+1) \inf_{m\in\mathcal{M}} \left\{d_2(s, S_m) \vee \eta_m\right\}^q, \quad \textit{for all } q \geq 1. \tag{3.9}$$

Proof: Since (3.9) is merely a version of (7.6) of Birgé (2006a) with $d = d_2$, we just have to show that Theorem 5 of this paper applies to our situation. It relies on Assumptions 1 and 3 of the paper. Assumption 3 follows from (3.7). As to Assumption 1 (with $a = n/(65\Gamma)$, $B = B' = 1$ and $\delta = 4d_2$, hence $\kappa = 4$), it is a consequence of our Theorem 2. The conditions (7.2) and (7.4) of Birgé (2006a) on η_m and D_m follow from (3.8). □

In the case of a single D-dimensional model $\overline{S} \subset \overline{\mathbb{L}}_\infty^\Gamma$ we get the following corollary:

Corollary 2 *Assume that we observe n i.i.d. random variables with unknown distribution P_s, $s \in \left(\overline{\mathbb{L}}_2, d_2\right)$ and that we have at disposal a D-dimensional model $\overline{S} \subset \overline{\mathbb{L}}_\infty^\Gamma$ for some given $\Gamma > 1$. One can build a T-estimator $\hat{s}$ such that, for all $s \in \overline{\mathbb{L}}_2$,*

$$\mathbb{E}_s\left[\|s-\hat{s}\|^2\right] \leq C\left[\inf_{t\in\overline{S}} d_2^2(s,t) + n^{-1}D\Gamma\right].$$

Proof: By Definition 1 and the remark following it, for each $\eta_0 > 0$, one can find an η_0-net $S_0 \subset \overline{S}$ for $\overline{S}$, hence $S_0 \subset \overline{\mathbb{L}}_\infty^\Gamma$, satisfying (3.7) with $D_0 = 25D/4$. Moreover $d(s, S_0) \leq \eta_0 + d\left(s, \overline{S}\right)$. Choosing $\eta_0^2 = 273 \times 25\Gamma D/4$, we may apply Theorem 3. The result then follows from (3.9) with $q = 2$. □

Theorem 3 applies in particular to the special situation of each model S_m being reduced to a single point $\{t_m\}$ so that we can take $D_m = 1/2$ for each m. We then get the following useful corollary.

Corollary 3 *Assume that we observe n i.i.d. random variables with unknown distribution P_s, $s \in \left(\overline{\mathbb{L}}_2, d_2\right)$ and that we have at disposal a countable subset $S = \{t_m\}_{m\in\mathcal{M}}$ of $\overline{\mathbb{L}}_\infty^\Gamma$ for some given $\Gamma > 1$. Let $\{\Delta_m\}_{m\in\mathcal{M}}$ be a family of weights such that $\Delta_m \geq 1/10$ for all $m \in \mathcal{M}$ and*

$$1 \leq \sum_{m\in\mathcal{M}} \exp[-\Delta_m] = \Sigma < +\infty. \tag{3.10}$$

We can build a T-estimator $\hat{s}$ such that, for all $s \in \overline{\mathbb{L}}_2$,

$$\mathbb{E}_s\left[d_2^q(s,\hat{s})\right] \le C_q\Sigma \inf_{m\in\mathcal{M}}\left\{d_2(s,t_m)\vee\sqrt{\Gamma\Delta_m/n}\right\}^q \quad \textit{for all } q\ge 1.$$

Proof: Let us set here $S_m=\{t_m\}$, $D_m=1/2$ and $\eta_m=37\sqrt{\Gamma\Delta_m/n}$ for $m\in\mathcal{M}$. One can then check that (3.7) and (3.8) are satified so that (3.9) holds. Our risk bound follows. □

At this stage, there are two main difficulties to apply Theorem 3 or Corollary 3. The first problem is to build suitable subsets S_m (or S) of $\overline{\mathbb{L}}_\infty^\Gamma$ from classical approximating sets (models), finite dimensional linear spaces for instance, that belong to $\mathbb{L}_2(\mu)$. We shall address this problem in the next section while we shall solve the second problem, namely choosing a convenient value for Γ from the data, in Section 5.

4 Model selection with uniformly bounded models

4.1 The projection operator onto $\overline{\mathbb{L}}_\infty^\Gamma$

Our first task is to define a projection operator π_Γ from $\mathbb{L}_2(\mu)$ onto $\overline{\mathbb{L}}_\infty^\Gamma$ ($\Gamma>1$) and to study its properties. In the sequel, we systematically identify a real number a with the function $a1\!\!1_{\mathcal{X}}$ for the sake of simplicity. The following proposition is the corrected version, by Yannick Baraud, of the initial mistaken result of the author.

Proposition 4 *For $t\in\mathbb{L}_2(\mu)$ and $1<\Gamma<+\infty$ we set $\pi_\Gamma(t)=[(t+\gamma)\vee 0]\wedge\Gamma$ where γ is defined by $\int[(t+\gamma)\vee 0]\wedge\Gamma\,d\mu=1$. Then π_Γ is the projection operator from $\mathbb{L}_2(\mu)$ onto the convex set $\overline{\mathbb{L}}_\infty^\Gamma$. Moreover, if $s\in\overline{\mathbb{L}}_2$ and $\Gamma>2$, then*

$$\|s-\pi_\Gamma(s)\|^2 \le \frac{\Gamma^2-\Gamma-1}{\Gamma(\Gamma-2)}Q_s(\Gamma) \quad \textit{with} \quad Q_s(z)=\int_{s>z}(s-z)^2\,d\mu. \tag{4.1}$$

Proof: First note that the existence of γ follows from the continuity and monotonicity of the mapping $z\mapsto\int[(t+z)\vee 0]\wedge\Gamma\,d\mu$ and that $\pi_\Gamma(t)\in\overline{\mathbb{L}}_\infty^\Gamma$. Since $\overline{\mathbb{L}}_\infty^\Gamma$ is a closed convex subset of a Hilbert space, the projection operator π onto $\overline{\mathbb{L}}_\infty^\Gamma$ exists and is characterized by the fact that

$$\langle t-\pi(t),u-\pi(t)\rangle\le 0 \quad \text{for all } u\in\overline{\mathbb{L}}_\infty^\Gamma. \tag{4.2}$$

Since $\int[u-\pi(t)]\,d\mu=0$ for $u\in\overline{\mathbb{L}}_\infty^\Gamma$, (4.2) implies that $\langle t+z-\pi(t),u-\pi(t)\rangle\le 0$ for $z\in\mathbb{R}$, hence $\pi(t)=\pi(t+z)$. Since $\pi_\Gamma(t)=\pi_\Gamma(t+z)$ as well, we may assume that $\int[t\vee 0]\wedge\Gamma\,d\mu=1$, hence $\pi_\Gamma(t)=[t\vee 0]\wedge\Gamma$ and $\pi_\Gamma(t)=t$ on the set $0\le t\le\Gamma$. Then, for $u\in\overline{\mathbb{L}}_\infty^\Gamma$,

$$\langle t-\pi_\Gamma(t),u-\pi_\Gamma(t)\rangle=\int_{t<0}tu\,d\mu+\int_{t>\Gamma}(t-\Gamma)(u-\Gamma)\,d\mu\le 0,$$

since $0\le u\le\Gamma$. This concludes the proof that $\pi=\pi_\Gamma$.

Let us now bound $\|s-\pi_\Gamma(s)\|$ when $s\in\overline{\mathbb{L}}_2$, setting $s=s\wedge\Gamma+v$ with $v=(s-\Gamma)\mathbb{1}_{s>\Gamma}$. Since there is nothing to prove when $\|s\|_\infty\le\Gamma$, we assume that $\int v\,d\mu>0$. By Cauchy-Schwarz Inequality,

$$\left(\int v\,d\mu\right)^2\le\mu(\{s>\Gamma\})\int v^2\,d\mu\le\Gamma^{-1}\|v\|^2. \tag{4.3}$$

Moreover, since $\int s\wedge\Gamma\,d\mu<1$, $\pi_\Gamma(s)=(s+\gamma)\wedge\Gamma$ with $0<\gamma\le 1$. Hence

$$\begin{aligned}1 &= \int[(s+\gamma)\wedge\Gamma]\,d\mu \ \ge\ \int(s\wedge\Gamma)\,d\mu+\gamma\mu(\{s\le\Gamma-\gamma\})\\ &\ge 1-\int v\,d\mu+\gamma\left(1-\frac{1}{\Gamma-\gamma}\right)\ >\ 1-\int v\,d\mu+\gamma\frac{\Gamma-2}{\Gamma-1}\end{aligned}$$

and $\gamma<(\Gamma-1)/(\Gamma-2)\int v\,d\mu$. Now, since $0\le\pi_\Gamma(s)-s\le\gamma$ for $s\le\Gamma$,

$$\begin{aligned}\|s-\pi_\Gamma(s)\|^2 &= \int_{s\le\Gamma}[\pi_\Gamma(s)-s]^2\,d\mu+\|v\|^2\ \le\ \gamma\int_{s\le\Gamma}[\pi_\Gamma(s)-s]\,d\mu+\|v\|^2\\ &< \frac{\Gamma-1}{\Gamma-2}\left(\int v\,d\mu\right)\int_{s>\Gamma}[s-\pi_\Gamma(s)]\,d\mu+\|v\|^2\\ &\le \frac{\Gamma-1}{\Gamma-2}\left(\int v\,d\mu\right)^2+\|v\|^2\ \le\ \left(1+\frac{\Gamma-1}{\Gamma(\Gamma-2)}\right)\|v\|^2,\end{aligned}$$

where we used (4.3). This concludes our proof. □

4.2 Selection for uniformly bounded countable sets

We consider here the situation mentioned in Corollary 3 but without the assumption that $S\subset\overline{\mathbb{L}}_\infty^\Gamma$. For $S=\{t_m\}_{m\in\mathcal{M}}$ an arbitrary countable subset of $\mathbb{L}^2(\mu)$ we may always replace it by its projection $\pi_\Gamma(S)$ onto $\overline{\mathbb{L}}_\infty^\Gamma$ and apply Corollary 3. The resulting risk bound involves

$$\begin{aligned}d_2\left(s,\pi_\Gamma(t_m)\right) &\le d_2\left(s,\pi_\Gamma(s)\right)+d_2\left(\pi_\Gamma(s),\pi_\Gamma(t_m)\right)\\ &\le \left(\frac{\Gamma^2-\Gamma-1}{\Gamma(\Gamma-2)}Q_s(\Gamma)\right)^{1/2}+d_2(s,t_m)\end{aligned}$$

by Proposition 4. We finally get:

Corollary 4 *Assume that we observe n i.i.d. random variables with unknown density $s\in\left(\overline{\mathbb{L}}_2,d_2\right)$ and that we have at disposal a countable subset $S=\{t_m\}_{m\in\mathcal{M}}$ of $\mathbb{L}^2(\mu)$ and a family of weights $\{\Delta_m\}_{m\in\mathcal{M}}$ such that $\Delta_m\ge 1/10$ for all $m\in\mathcal{M}$ and (3.10) holds. Given $\Gamma\ge 3$ we can build a T-estimator $\hat{s}^\Gamma$ with values in $\pi_\Gamma(S)$ such that, for all $s\in\overline{\mathbb{L}}_2$,*

$$\mathbb{E}_s\left[\left\|s-\hat{s}^\Gamma\right\|^q\right]\le C_q\Sigma\inf_{m\in\mathcal{M}}\left\{\left[d_2(s,t_m)+\sqrt{Q_s(\Gamma)}\right]\vee\sqrt{\Gamma\Delta_m/n}\right\}^q\quad \textit{for } q\ge 1.$$

4.3 Selection with uniformly bounded models

Typical models $\overline{S}$ for density estimation in $\mathbb{L}_2(\mu)$ are finite-dimensional linear spaces which are not subsets of $\overline{\mathbb{L}}_\infty^\Gamma$ but merely spaces of functions with nice approximation properties. To apply Theorem 3 we have to replace them by discrete subsets of $\overline{\mathbb{L}}_\infty^\Gamma$ that satisfy (3.7). Unfortunately, they cannot simply be derived by a discretization of $\overline{S}$ followed by a projection π_Γ or a discretization of $\pi_\Gamma\left(\overline{S}\right)$. A more complicated construction is required to preserve both the metric and approximation properties of $\overline{S}$. It is provided by the following preliminary result.

Proposition 5 *Let $\overline{S}$ be a subset of $\mathbb{L}_2(\mu)$ with metric dimension bounded by D. For $\Gamma > 2$ and $\eta > 0$, one can find a discrete subset S' of $\overline{\mathbb{L}}_\infty^\Gamma$ with the following properties:*

$$|S' \cap \mathcal{B}_{d_2}(t, x\eta)| \leq \exp\left[9Dx^2\right] \quad \textit{for all } x \geq 2 \textit{ and } t \in \mathbb{L}_2(\mu); \tag{4.4}$$

for any $s \in \overline{\mathbb{L}}_2$, one can find some $s' \in S'$ such that

$$\|s - s'\| \leq 3.1\left[\eta + \inf_{t\in\overline{S}} \|s - t\|\right] + 4.1\left(\frac{\Gamma^2 - \Gamma - 1}{\Gamma(\Gamma - 2)} Q_s(\Gamma)\right)^{1/2}. \tag{4.5}$$

Proof: According to Definition 1, we choose some η-net S_η for $\overline{S}$ such that (2.2) holds for all $t \in \mathbb{L}_2(\mu)$. Since, by Proposition 4, the operator π_Γ from $\mathbb{L}_2(\mu)$ to $\overline{\mathbb{L}}_\infty^\Gamma$ satisfies $\|u - \pi_\Gamma(t)\| \leq \|u - t\|$ for all $u \in \overline{\mathbb{L}}_\infty^\Gamma$, we may apply Proposition 12 of Birgé (2006a) with $M' = \mathbb{L}_2(\mu)$, $d = d_2$, $M_0 = \overline{\mathbb{L}}_\infty^\Gamma$, $T = S_\eta$, $\overline{\pi} = \pi_\Gamma$ and $\lambda = 1$. It shows that one can find a subset S' of $\pi_\Gamma(S_\eta)$ such that (4.4) holds and $d_2(u, S') \leq 3.1 d_2(u, S_\eta)$ for all $u \in \overline{\mathbb{L}}_\infty^\Gamma$. If s is an arbitrary element of $\overline{\mathbb{L}}_2$, then

$$d_2\left(\pi_\Gamma(s), S'\right) \leq 3.1 d_2\left(\pi_\Gamma(s), S_\eta\right) \leq 3.1\left[d_2\left(\pi_\Gamma(s), s\right) + d_2\left(s, \overline{S}\right) + \eta\right],$$

hence

$$d_2\left(s, S'\right) \leq 3.1\left[d_2\left(s, \overline{S}\right) + \eta\right] + 4.1 d_2\left(\pi_\Gamma(s), s\right). \tag{4.6}$$

The conclusion follows from Proposition 4. □

We are now in a position to derive our main result about bounded model selection. We start with a countable collection $\{\overline{S}_m, m \in \mathcal{M}\}$ of models in $\mathbb{L}_2(\mu)$ with metric dimensions bounded respectively by $\overline{D}_m \geq 1/2$ and a family of weights Δ_m satisfying (3.10). We fix some $\Gamma \geq 3$ and, for each $m \in \mathcal{M}$, we set

$$\eta_m = \left[\left(50\sqrt{\overline{D}_m}\right) \vee \left(37\sqrt{\Delta_m}\right)\right]\sqrt{\Gamma/n}.$$

By Proposition 5 (with $\eta = \eta_m$), each $\overline{S}_m$ gives rise to a subset S_m^Γ which satisfies (3.7) with $D_m = 9\overline{D}_m$. It follows from our choice of η_m that (3.8) is also satisfied so that we may apply Theorem 3 to the family of sets $\left\{S_m^\Gamma, m \in \mathcal{M}\right\}$. This results in a T-estimator $\hat{s}^\Gamma$ such that, for all $s \in \overline{\mathbb{L}}_2$,

$$\mathbb{E}_s\left[d_2^q\left(s, \hat{s}^\Gamma\right)\right] \leq C_q \Sigma \inf_{m\in\mathcal{M}}\left\{d_2\left(s, S_m^\Gamma\right) \vee \eta_m\right\}^q \quad \text{for } q \geq 1.$$

We also derive from Proposition 5 that

$$d_2\big(s,S_m^\Gamma\big) \le 3.1\left[\eta_m + \inf_{t\in\overline{S}_m}\|s-t\|\right] + 4.1\sqrt{(5/3)Q_s(\Gamma)}.$$

Putting the bounds together and rearranging the terms leads to the following theorem.

Theorem 4 *Given a countable collection $\{\overline{S}_m, m\in\mathcal{M}\}$ of models in $\mathbb{L}_2(\mu)$ with metric dimensions bounded respectively by $\overline{D}_m \ge 1/2$ and a family of weights Δ_m satisfying (3.10), one can build, for each $\Gamma\ge 3$, a T-estimator $\hat{s}^\Gamma$ which satisfies, for all $s\in\overline{\mathbb{L}}_2$ and $q\ge 1$,*

$$\mathbb{E}_s\left[\left\|s-\hat{s}^\Gamma\right\|^q\right] \le C_q\Sigma\left[\inf_{m\in\mathcal{M}}\left\{d_2\big(s,\overline{S}_m\big)+\sqrt{\frac{\Gamma\left(\overline{D}_m\vee\Delta_m\right)}{n}}\right\}+\sqrt{Q_s(\Gamma)}\right]^q, \tag{4.7}$$

with Q_s given by (4.1) and C_q some constant depending only on q. If $\|s\|_\infty\le\Gamma$, then

$$\mathbb{E}_s\left[\left\|s-\hat{s}^\Gamma\right\|^2\right] \le C\Sigma\inf_{m\in\mathcal{M}}\left\{d_2^2\big(s,\overline{S}_m\big)+n^{-1}\Gamma\left(\overline{D}_m\vee\Delta_m\right)\right\}. \tag{4.8}$$

5 A selection theorem

Now that we are able to build models which are bounded by Γ for each $\Gamma\ge 3$ and to select one of these models, which results in an estimator $\hat{s}^\Gamma$, we need a way to choose Γ from the data in order to optimize the bound in (4.7). The idea is to use one half of our sample to build a sequence of estimators $\hat{s}^{2^i}$ and select a convenient value of i from the other half of our sample. This requires to select an element from a sequence of densities which is not uniformly bounded.

5.1 A preliminary selection result

We start with a general selection result, to be proved in Section 7.3, that we state for an arbitrary statistical framework since it may apply to other situations than density estimation from an i.i.d. sample. We observe some random object $\boldsymbol{X}$ with distribution P_s on $\mathcal{X}$ where s belongs to a metric space M (carrying a distance d) which indexes a family $\mathcal{P}=\{P_t, t\in M\}$ of probabilities on $\mathcal{X}$.

Theorem 5 *Let $(t_p)_{p\ge1}$ be a sequence in M such that the following assumption holds: for all pairs (n,p) with $1\le n<p$ and all $x\in\mathbb{R}$, one can find a test $\psi_{t_n,t_p,x}$ based on the observation $\boldsymbol{X}$ and satisfying*

$$\sup_{\{s\in M\,|\,d(s,t_n)\le d(t_n,t_p)/4\}}\mathbb{P}_s[\psi_{t_n,t_p,x}(\boldsymbol{X})=t_p]\le B\exp\left[-a2^{-p}d^2(t_n,t_p)-x\right]; \tag{5.1}$$

$$\sup_{\{s\in M\,|\,d(s,t_p)\le d(t_n,t_p)/4\}}\mathbb{P}_s[\psi_{t_n,t_p,x}(\boldsymbol{X})=t_n]\le B\exp\left[-a2^{-p}d^2(t_n,t_p)+x\right]; \tag{5.2}$$

with positive constants a and B independent of n,p and x. For each $A\ge1$, one can design an estimator $\hat{s}_A$ such that, for all $s\in M$,

$$\mathbb{E}_s\left[d^q\left(\hat{s}_A,s\right)\right]\le BC(A,q)\inf_{p\ge1}\left[d(s,t_p)\vee\sqrt{a^{-1}p2^p}\right]^q \quad \textit{for } 1\le q<2A/\log 2. \tag{5.3}$$

This general result applies to our specific framework of density estimation based on an observation $\boldsymbol{X}$ with distribution P_s, $s \in \overline{\mathbb{L}}_2$, provided that the sequence $(t_p)_{p\geq 1}$ be suitably chosen. We shall simply assume here that $t_p \in \overline{\mathbb{L}}_2$ with $\|t_p\|_\infty \leq 2^{p+1}$ for each $p \geq 1$. This implies that, for $1 \leq i < j$, t_i and t_j belong to $\overline{\mathbb{L}}_\infty^{2^{j+1}}$ so that Theorem 2 applies with $\boldsymbol{X}$ replaced by the randomized sample $\boldsymbol{X}'$ and the assumption of Theorem 5 is therefore satisfied with $d = d_2$, $B = 1$ and $a = n/65$, leading to the following corollary.

Corollary 5 *Let $(t_i)_{i\geq 1}$ be a sequence of densities such that $t_i \in \overline{\mathbb{L}}_\infty^{2^{i+1}}$ for each i, $A \geq 1$ and $\boldsymbol{X}$ be an n-sample with density $s \in \overline{\mathbb{L}}_2$. One can design an estimator $\hat{s}_A$ such that*

$$\mathbb{E}_s\left[d_2^q(\hat{s}_A, s)\right] \leq C(A,q) \inf_{i\geq 1}\left[d_2(s,t_i) \vee \sqrt{n^{-1}i2^i}\right]^q \quad \text{for } 1 \leq q < 2A/\log 2.$$

5.2 General model selection in $\overline{\mathbb{L}}_2$

We now consider the general situation where we observe $n = 2n'$ i.i.d. random variables $X_1, \ldots, X_n$ with an unknown density $s \in \overline{\mathbb{L}}_2$, not necessarily bounded, and have at disposal a countable collection $\{\overline{S}_m, m \in \mathcal{M}\}$ of models in $\mathbb{L}_2(\mu)$ with metric dimensions bounded respectively by $\overline{D}_m \geq 1/2$ and a family of weights Δ_m which satisfy (3.10). We split our sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ into two subsamples $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ of size n'. With the sample $\boldsymbol{X}_1$ we build the T-estimators $\hat{s}_i(\boldsymbol{X}_1) = \hat{s}^{2^{i+1}}(\boldsymbol{X}_1)$, $i \geq 1$ which are provided by Theorem 3. It then follows from (4.7) that each such estimator satisfies, for $q \geq 1$,

$$\begin{aligned}
&\mathbb{E}_s\left[\|s - \hat{s}_i(\boldsymbol{X}_1)\|^q\right] \\
&\quad \leq \quad C_q\Sigma\left\{\inf_{m\in\mathcal{M}}\left[d_2\left(s,\overline{S}_m\right) + \left(\frac{2^i\left(\overline{D}_m \vee \Delta_m\right)}{n}\right)^{1/2}\right] + \sqrt{Q_s(2^{i+1})}\right\}^q,
\end{aligned}$$

with Q_s given by (4.1). We now work conditionally on $\boldsymbol{X}_1$, fix a convenient value of $A \geq 1$ (for instance $A = 1$ if we just want to bound the quadratic risk) and use the second half of the sample $\boldsymbol{X}_2$ to select one estimator among the previous family according to the procedure described in Section 5.1. By Corollary 5 this results in a new estimator $\tilde{s}_A(\boldsymbol{X})$ which satisfies

$$\mathbb{E}_s\left[d_2^q(\tilde{s}_A(\boldsymbol{X}), s) \,\middle|\, \boldsymbol{X}_1\right] \leq C(A,q) \inf_{i\geq 1}\left[d_2(s, \hat{s}_i(\boldsymbol{X}_1)) \vee \sqrt{n^{-1}i2^i}\right]^q,$$

provided that $q < 2A/\log 2$. Integrating with respect to $\boldsymbol{X}_1$ and using our previous risk bound gives

$$\begin{aligned}
&\mathbb{E}_s\left[\|s - \tilde{s}_A(\boldsymbol{X})\|^q\right] \\
&\quad \leq \quad C(A,q)\inf_{i\geq 1}\left\{\mathbb{E}_s\left[\|s - \hat{s}_i(\boldsymbol{X}_1)\|^q\right] + \left(n^{-1}i2^i\right)^{q/2}\right\} \\
&\quad \leq \quad C'(A,q)\Sigma\inf_{i\geq 1}\left\{\inf_{m\in\mathcal{M}}\left[d_2^q\left(s,\overline{S}_m\right) + \left(\frac{2^i\left(\overline{D}_m \vee \Delta_m \vee i\right)}{n}\right)^{q/2}\right] + \left[Q_s(2^{i+1})\right]^{q/2}\right\}.
\end{aligned}$$

For $2^i \leq z < 2^{i+1}$, $\log z \geq i \log 2$ and $Q_s(z) \geq Q_s(2^{i+1})$ since Q_s is nonincreasing. Modifying accordingly the constants in our bounds, we get the main result of this paper which provides adaptation to both the models and the truncation constant.

Theorem 6 *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ with $n \geq 2$ be an i.i.d. sample with unknown density $s \in \overline{\mathbb{L}}_2$ and $\{\overline{S}_m, m \in \mathcal{M}\}$ be a countable collection of models in $\mathbb{L}_2(\mu)$ with metric dimensions bounded respectively by $\overline{D}_m \geq 1/2$. Let $\{\Delta_m, m \in \mathcal{M}\}$ be a family of weights which satisfy (3.10) and $Q_s(z)$ be given by (4.1). For each $A \geq 1$, there exists an estimator $\tilde{s}_A(\boldsymbol{X})$ such that, whatever $s \in \overline{\mathbb{L}}_2$ and $1 \leq q < (2A/\log 2)$,*

$$\begin{aligned} &\mathbb{E}_s\left[\|s - \tilde{s}_A(\boldsymbol{X})\|^q\right] \\ &\quad \leq C(A,q)\Sigma \inf_{z \geq 2} \inf_{m \in \mathcal{M}} \left[d_2^q\left(s, \overline{S}_m\right) + \left(\frac{z\left(\overline{D}_m \vee \Delta_m \vee \log z\right)}{n} \right)^{q/2} + \left[Q_s(z)\right]^{q/2} \right]. \end{aligned} \tag{5.4}$$

In particular, for $\tilde{s} = \tilde{s}_1$ and $s \in \overline{\mathbb{L}}_\infty(\mu)$,

$$\mathbb{E}_s\left[\|s - \tilde{s}(\boldsymbol{X})\|^2\right] \leq C\Sigma \inf_{m \in \mathcal{M}} \left[d_2^2\big(s, \overline{S}_m\big) + n^{-1}\|s\|_\infty \left(\overline{D}_m \vee \Delta_m \vee \log \|s\|_\infty\right)\right]. \tag{5.5}$$

5.3 Some remarks

We see that (5.4) is a generalization of (4.7) and (5.5) of (4.8) at the modest price of the extra $\log z$ (or $\log \|s\|_\infty$). We do not know whether this $\log z$ is necessary or not but, in a typical model selection problem, when s belongs to $\overline{\mathbb{L}}_\infty(\mu)$ but not to $\cup_{m \in \mathcal{M}} \overline{S}_m$, the optimal value of $\overline{D}_m$ goes to $+\infty$ with n, so that, for this optimal value, asymptotically $\overline{D}_m \vee \Delta_m \vee \log \|s\|_\infty = \overline{D}_m \vee \Delta_m$.

Up to constants depending on $\|s\|_\infty$, (5.5) is the exact analogue of (2.5) which shows that, when $s \in \overline{\mathbb{L}}_\infty(\mu)$, all the results about model selection obtained for the Hellinger distance can be translated in terms of the $\mathbb{L}_2$-distance.

Note that Theorem 6 applies to a single model $\overline{S}$ with metric dimension bounded by $\overline{D}$, in which case one can use a weight $\Delta = 1/2 \leq \overline{D}$ which results, if $A = 1$, in the risk bound

$$\mathbb{E}_s\left[\|s - \tilde{s}(\boldsymbol{X})\|^2\right] \leq C\left[d_2^2\left(s, \overline{S}\right) + \inf_{z \geq 2}\left\{\frac{z\left(\overline{D} \vee \log z\right)}{n} + Q_s(z)\right\}\right], \tag{5.6}$$

and, if $s \in \overline{\mathbb{L}}_\infty(\mu)$,

$$\mathbb{E}_s\left[\|s - \tilde{s}(\boldsymbol{X})\|^2\right] \leq C\left[d_2^2\big(s, \overline{S}\big) + n^{-1}\|s\|_\infty \left(\overline{D} \vee \log \|s\|_\infty\right)\right]. \tag{5.7}$$

Apart from the extra $\log \|s\|_\infty$, which is harmless when it is smaller than $\overline{D}$, we recover what we expected, namely the bound (2.8).

Even if $s \in \overline{\mathbb{L}}_\infty(\mu)$ the bound (5.4) may be much better than (5.5). This is actually already visible with one single model, comparing (5.6) with (5.7). It is indeed easy to find an example of a very spiky density s for which (5.6) is much better than (5.7) or the classical bound (2.6) obtained for projection estimators. Of course, this is just a comparison of universal bounds, not of the real risk of estimators for a given s.

More surprising is the fact that our estimator can actually dominate a histogram based on the same model, although our counter-example is rather caricatural and more an advertising against the use of the $\mathbb{L}_2$-loss than against the use of histogram estimators. Let us consider a partition $\mathcal{I}$ of $[0,1]$ into $2D$ intervals I_j, $1 \le j \le 2D$ with the integer D satisfying $2 \le D \le n$ and fix some $\gamma \ge 10$. We then set $\alpha = \left(\gamma^2 n\right)^{-1}$. For $1 \le j \le D$, the intervals I_{2j-1} have length α while the intervals I_{2j} have length β with $D(\alpha+\beta) = 1$. We denote by $\overline{S}$ the $2D$-dimensional linear space spanned by the indicator functions of the I_j. It is a model with metric dimension bounded by D. We assume that the underlying density s with respect to Lebesgue measure belongs to $\overline{S}$ and is defined as

$$s = p\alpha^{-1}\sum_{j=1}^{D}1\!\!1_{I_{2j-1}} + q\beta^{-1}\sum_{j=1}^{D}1\!\!1_{I_{2j}} \quad \text{with } p = \gamma\alpha \quad \text{and} \quad D(p+q) = 1,$$

so that $\beta > q$ since $\alpha < p$. We consider two estimators of s derived from the same model $\overline{S}$: the histogram $\hat{s}_{\mathcal{I}}$ based on the partition $\mathcal{I}$ and the estimator $\tilde{s}$ based on $\overline{S}$ and provided by Theorem 6. According to (1.2) the risk of $\hat{s}_{\mathcal{I}}$ is

$$Dn^{-1}\left[\alpha^{-1}p(1-p) + \beta^{-1}q(1-q)\right] \ge 0.9Dn^{-1}\alpha^{-1}p = 0.9D\gamma n^{-1},$$

since $p \le 1/10$. The risk of $\tilde{s}$ can be bounded by (5.4) with $z = 4$ which gives

$$\mathbb{E}_s\left[\|s - \tilde{s}(\boldsymbol{X})\|^2\right] \le C\left[4Dn^{-1} + D\int_{I_1}(p/\alpha)^2\,d\mu\right] = CD\left[4n^{-1} + p^2\alpha^{-1}\right] = 5CDn^{-1}.$$

For large enough values of γ our estimator is better than the histogram. The problem actually comes from the observations falling in some of the intervals I_{2j-1} which will lead to a very bad estimation of s on those intervals. Note that this fact will happen with a small probability since $Dp = D(\gamma n)^{-1} \le \gamma^{-1}$. Nevertheless, this event of small probability is important enough to lead to a large risk when we use the $\mathbb{L}_2$-loss.

6 Some applications

6.1 Aggregation of preliminary estimators

Theorem 6 applies in particular to the problem of aggregating preliminary estimators, built from an independent sample, either by selecting one of them or by combining them linearly.

6.1.1 Aggregation by selection

Let us begin with the problem, that we already considered in Section 4.2, of selecting a point among a countable family $\{t_m, m \in \mathcal{M}\}$. Typically, as in Rigollet (2006), the t_m are preliminary estimators based on an independent sample (derived by sample splitting if necessary) and we want to choose the best one in the family. This is a situation for which one can choose $\overline{D}_m = 1/2$ and $A = 1$ which leads to the following corollary

Corollary 6 *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ with $n \ge 2$ be an i.i.d. sample with unknown density $s \in \overline{\mathbb{L}}_2$ and $\{t_m, m \in \mathcal{M}\}$ be a countable collection of points in $\mathbb{L}_2(\mu)$. Let*

$\{\Delta_m, m \in \mathcal{M}\}$ be a family of weights which satisfy (3.10) and $Q_s(z)$ be given by (4.1). There exists an estimator $\tilde{s}(\boldsymbol{X})$ such that, whatever $s \in \overline{\mathbb{L}}_2$,

$$\mathbb{E}_s\left[\|s-\tilde{s}(\boldsymbol{X})\|^2\right] \leq C\Sigma \inf_{z\geq 2}\left\{\inf_{m\in\mathcal{M}}\left[d_2^2(s,t_m)+\frac{z(\Delta_m \vee \log z)}{n}\right]+Q_s(z)\right\}.$$

6.1.2 Linear aggregation

Rigollet and Tsybakov (2007) have considered the problem of linear aggregation. Given a finite set $\{t_1,\ldots,t_N\}$ of preliminary estimators of s, they use the observations to build a linear combination of the t_j in order to get a new and potentially better estimator of s. For $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_N) \in \mathbb{R}^N$, let us set $t_{\boldsymbol{\lambda}} = \sum_{j=1}^N \lambda_j t_j$. Rigollet and Tsybakov build a selector $\widehat{\boldsymbol{\lambda}}(X_1,\ldots,X_n)$ such that the corresponding estimator $\widehat{s}(\boldsymbol{X}) = t_{\widehat{\boldsymbol{\lambda}}}$ satisfies, for all $s \in \overline{\mathbb{L}}_\infty$,

$$\mathbb{E}_s\left[\|s-\widehat{s}(\boldsymbol{X})\|^2\right] \leq \inf_{\boldsymbol{\lambda}\in\mathbb{R}^N} d_2^2\left(s,t_{\boldsymbol{\lambda}}\right)+n^{-1}\|s\|_\infty N. \tag{6.1}$$

Unfortunately, this bound, which is shown to be sharp for such an estimator, can be really poor, as compared to the minimal risk $\inf_{1\leq j\leq N} d_2^2(s,t_j)$ of the preliminary estimators when one of those is already quite good and $n^{-1}\|s\|_\infty N$ is large, which is likely to happen when N is quite large. Moreover, this result tells nothing when $s \notin \overline{\mathbb{L}}_\infty$. In Birgé (2006a, Section 9.3) we proposed an alternative way of selecting a linear combination of the t_j based on T-estimators. In the particular situation of densities belonging to $\overline{\mathbb{L}}_2$, we proceed as follows: we choose for $\mathcal{M}$ the collection of all nonvoid subsets m of $\{1,\ldots,N\}$ and, for $m \in \mathcal{M}$, we take for $\overline{S}_m$ the linear span of the t_j with $j \in m$ so that the dimension of $\overline{S}_m$ is bounded by $|m|$ and its metric dimension $\overline{D}_m$ by $|m|/2$. Since the number of elements of $\mathcal{M}$ with cardinality j is $\begin{pmatrix} N \\ j \end{pmatrix} < (eN/j)^j$, we may set $\Delta_m = |m|[2+\log(N/|m|)]$ so that (3.10) is satisfied with $\Sigma < 1$. An application of Theorem 6 leads to the following corollary.

Corollary 7 *Let $\boldsymbol{X} = (X_1,\ldots,X_n)$ with $n \geq 2$ be an i.i.d. sample with unknown density $s \in \overline{\mathbb{L}}_2$ and $\{t_1,\ldots,t_N\}$ be a finite set of points in $\mathbb{L}_2(\mu)$. Let $\mathcal{M}$ be the collection of all nonvoid subsets m of $\{1,\ldots,N\}$ and, for $m \in \mathcal{M}$,*

$$\Lambda_m = \left\{\boldsymbol{\lambda} \in \mathbb{R}^N \mid \lambda_j = 0 \text{ for } j \notin m\right\}.$$

For each $A \geq 1$, there exists an estimator $\tilde{s}_A(\boldsymbol{X})$ such that, whatever $s \in \overline{\mathbb{L}}_2$ and $1 \leq q < (2A/\log 2)$,

$$\mathbb{E}_s\left[\|s-\tilde{s}_A(\boldsymbol{X})\|^q\right] \leq C(A,q)\inf_{z\geq 2}\inf_{m\in\mathcal{M}} R(q,s,z,m),$$

where

$$R(q,s,z,m) = \inf_{\boldsymbol{\lambda}\in\Lambda_m} d_2^q\left(s,t_{\boldsymbol{\lambda}}\right)+\left(\frac{z\left[|m|\left(1+\log(N/|m|)\right)\vee\log z\right]}{n}\right)^{q/2}+[Q_s(z)]^{q/2}$$

and $Q_s(z)$ is given by (4.1).

There are many differences between this bound and (6.1), apart from the nasty constant $C(A,q)$. Firstly, it applies to densities s that do not belong to $\overline{\mathbb{L}}_\infty$ and handles the case of $q>2$ for a convenient choice of A. Also, when $s\in\overline{\mathbb{L}}_\infty$ and one of the preliminary estimators is already close to s, it may very well happen, when N is large, that

$$R\left(2,s,\|s\|_\infty,m\right)\le\inf_{\boldsymbol{\lambda}\in\Lambda_m}d_2^2\left(s,t_{\boldsymbol{\lambda}}\right)+n^{-1}\|s\|_\infty\left[|m|\left(1+\log(N/|m|)\right)\vee\log\|s\|_\infty\right]$$

be much smaller than the right-hand side of (6.1) for some m of small cardinality.

6.2 Selection of projection estimators

In this section, we assume that $s\in\overline{\mathbb{L}}_\infty(\mu)$. This assumption is not needed for the design of the estimator but only to derive suitable risk bounds. We have at hand a countable family $\left\{\overline{S}_m,m\in\mathcal{M}\right\}$ of linear subspaces of $\mathbb{L}_2(\mu)$ with respective dimensions D_m and we choose corresponding weights Δ_m satisfying (3.10). For each m, we consider the projection estimator $\hat{s}_m$ defined in Section 2.2. Each such estimator has a risk bounded by (2.6), i.e.

$$\mathbb{E}_s\left[\|\hat{s}_m-s\|^2\right]\le\|\overline{s}_m-s\|^2+n^{-1}D_m\|s\|_\infty,$$

where $\overline{s}_m$ denotes the orthogonal projection of s onto $\overline{S}_m$. If we apply Corollary 6 to this family of estimators, we get an estimator $\tilde{s}(\boldsymbol{X})$ satisfying, for all $s\in\overline{\mathbb{L}}_\infty$,

$$\mathbb{E}_s\left[\|s-\tilde{s}(\boldsymbol{X})\|^2\right]\le C\Sigma\inf_{m\in\mathcal{M}}\left[\|\overline{s}_m-s\|^2+n^{-1}\|s\|_\infty\left(D_m\vee\Delta_m\vee\log\|s\|_\infty\right)\right].$$

With this bound at hand, we can now return to the problem we considered in Section 1.1, starting with an arbitrary countable family $\{\mathcal{I}_m,m\in\mathcal{M}\}$ of finite partitions of $\mathcal{X}$ and weights Δ_m satisfying (3.10). To each partition $\mathcal{I}_m$ we associate the linear space $\overline{S}_m$ of piecewise constant fonctions of the form $\sum_{I\in\mathcal{I}_m}\beta_I\mathbb{1}_I$. The dimension of this linear space is the cardinality of $\mathcal{I}_m$ and its metric dimension is bounded by $|\mathcal{I}_m|/2$. If we know that $s\in\overline{\mathbb{L}}_\infty(\mu)$, we can proceed as we just explained, building the family of histograms $\hat{s}_{\mathcal{I}_m}(\boldsymbol{X}_1)$ corresponding to our partitions and using Corollary 6 to get

$$\mathbb{E}_s\left[\|s-\tilde{s}(\boldsymbol{X})\|^2\right]\le C\Sigma\inf_{m\in\mathcal{M}}\left[\|\overline{s}_{\mathcal{I}_m}-s\|^2+n^{-1}\|s\|_\infty\left(|\mathcal{I}_m|\vee\Delta_m\vee\log\|s\|_\infty\right)\right],\tag{6.2}$$

which should be compared with (1.3). Apart from the unavoidable complexity term Δ_m due to model selection, we have only lost (up to the universal constant C) the replacement of $|\mathcal{I}_m|$ by $|\mathcal{I}_m|\vee\log\|s\|_\infty$. Examples of families of partitions that satisfy (3.10) are given in Section 9 of Birgé (2006a).

In the general case of $s\in\overline{\mathbb{L}}_2(\mu)$, we may apply Theorem 6 to the family of linear models $\left\{\overline{S}_m,m\in\mathcal{M}\right\}$ derived from these partitions, getting an estimator $\tilde{s}$ with a risk satisfying

$$\mathbb{E}_s\left[\|s-\tilde{s}(\boldsymbol{X})\|^2\right]\le C\Sigma\inf_{z\ge2}\left\{\inf_{m\in\mathcal{M}}\left[\|\overline{s}_{\mathcal{I}_m}-s\|^2+\frac{z(|\mathcal{I}_m|\vee\Delta_m\vee\log z)}{n}\right]+Q_s(z)\right\}.$$

6.3 A comparison with Gaussian model selection

A benchmark for model selection in general is the particular (simpler) situation of model selection for the so-called *white noise framework* in which we observe a Gaussian process $\boldsymbol{X} = \{X_z, z \in [0,1]\}$ with $X_z = \int_0^z s(x)\,dx + \sigma W_z$, where s is an unknown element of $\mathbb{L}_2([0,1],dx)$, $\sigma > 0$ a known parameter and W_z a Wiener process. For such a problem, an analogue of Theorem 1 has been proved in Birgé (2006a), namely

Theorem 7 *Let $\boldsymbol{X}$ be the Gaussian process given by*

$$X_z = \int_0^z s(x)\,dx + n^{-1/2}W_z, \quad 0 \le z \le 1,$$

where s is an unknown element of $\mathbb{L}_2([0,1],dx)$ to be estimated and W_z a Wiener process. Let $\{\overline{S}_m, m \in \mathcal{M}\}$ be a countable collection of models in $\mathbb{L}_2(\mu)$ with metric dimensions bounded respectively by $\overline{D}_m \ge 1/2$. Let $\{\Delta_m, m \in \mathcal{M}\}$ be a family of weights which satisfy (3.10). There exists an estimator $\tilde{s}(\boldsymbol{X})$ such that, whatever $s \in \mathbb{L}_2([0,1],dx)$,

$$\mathbb{E}_s\left[\|s - \tilde{s}(\boldsymbol{X})\|^2\right] \le C \inf_{m\in\mathcal{M}}\left[d_2^2\left(s,\overline{S}_m\right) + n^{-1}\left(\overline{D}_m \vee \Delta_m\right)\right].$$

Comparing this bound with (5.5) shows that, when $s \in \overline{\mathbb{L}}_\infty(\mu)$, we get a similar risk bound for estimating the density s from n i.i.d. random variables, apart from an additional factor depending on $\|s\|_\infty$. Similar analogies are valid with bounds obtained for estimating densities with squared Hellinger loss or for estimating the intensity of a Poisson process as shown in Birgé (2006a and 2007). Therefore, all the many examples that have been treated in these papers could be transferred to the case of density estimation with $\mathbb{L}_2$-loss with minor modifications due to the appearence of $\|s\|_\infty$ in the bounds. We leave all these translations as exercices for the concerned reader.

6.4 Estimation in Besov spaces

The Besov space $B^\alpha_{p,\infty}([0,1])$ with $\alpha, p > 0$ is defined in DeVore and Lorentz (1993) and it is known that a necessary and sufficient condition for $B^\alpha_{p,\infty}([0,1]) \subset \mathbb{L}_2([0,1],dx)$ is $\delta = \alpha + 1/2 - 1/p > 0$, which we shall assume in the sequel. The problem of estimating adaptively densities that belong to some Besov space $B^\alpha_{p,\infty}([0,1])$ with unknown values of α and p has been solved for a long time when $\alpha > 1/p$ which is a necessary and sufficient condition for $B^\alpha_{p,\infty}([0,1]) \subset \mathbb{L}_\infty([0,1],dx)$. See for instance Donoho, Johnstone, Kerkyacharian and Picard (1996), Delyon and Juditsky (1996) or Birgé and Massart (1997). It can be treated in the usual way (with an estimation of $\|s\|_\infty$) leading to the minimax rate of convergence $n^{-2\alpha/(2\alpha+1)}$ for the quadratic risk.

6.4.1 Wavelet expansions

It is known from analysis that functions $s \in \mathbb{L}_2\left([0,1],dx\right)$ can be represented by their expansion with respect to some orthonormal wavelet basis $\{\varphi_{j,k}, j \ge -1, k \in \Lambda(j)\}$

with $|\Lambda(-1)| \leq K$ and $2^j \leq |\Lambda(j)| \leq K2^j$ for all $j \geq 0$. Such a wavelet basis satisfies

$$\left\| \sum_{k\in\Lambda(j)} \varphi_{j,k} \right\|_\infty \leq K' 2^{j/2} \quad \text{for } j \geq -1, \tag{6.3}$$

and we can write

$$s = \sum_{j=-1}^{\infty} \sum_{k\in\Lambda(j)} \beta_{j,k}\varphi_{j,k}, \quad \text{with} \quad \beta_{j,k} = \int \varphi_{j,k}(x)s(x)\,dx. \tag{6.4}$$

Moreover, for a convenient choice of the wavelet basis (depending on α), the fact that s belongs to the Besov space $B^\alpha_{p,\infty}([0,1])$ is equivalent to

$$\sup_{j\geq 0} 2^{j(\alpha+1/2-1/p)} \left(\sum_{k\in\Lambda(j)} |\beta_{j,k}|^p \right)^{1/p} = |s|_{\alpha,p,\infty} < +\infty, \tag{6.5}$$

where $|s|_{\alpha,p,\infty} < +\infty$ is equivalent to the Besov semi-norm $|s|^\alpha_p$.

Moreover, it follows from Birgé and Massart (1997 and 2000), as summarized in Birgé (2006a, Proposition 13), that, given the integer r, one can find a wavelet basis (depending on r) and a universal family of linear models $\{\overline{S}_m, m \in \mathcal{M} = \cup_{J\geq 0}\mathcal{M}_J\}$ with respective dimensions $\overline{D}_m$, and weights $\{\Delta_m, m \in \mathcal{M}\}$ satisfying (3.10), with the following properties. Each $\overline{S}_m$ is the linear span of $\{\varphi_{-1,k}, k \in \Lambda(-1)\} \cup \{\varphi_{j,k}, (j,k) \in m\}$ with $m \subset \cup_{j\geq 0}\Lambda(j)$; $\overline{D}_m \vee \Delta_m \leq c2^J$ for $m \in \mathcal{M}_J$ and

$$\inf_{m\in\mathcal{M}_J} \inf_{t\in\overline{S}_m} \|s-t\| \leq C(\alpha,p)2^{-J\alpha}|s|_{\alpha,p,\infty} \quad \text{for } s \in B^\alpha_{p,\infty}([0,1]),\ \alpha < r. \tag{6.6}$$

6.4.2 The bounded case

Actually, only the assumption that $s \in B^\alpha_{p,\infty}([0,1]) \cap \mathbb{L}_\infty(\mu)$, rather than $\alpha > 1/p$, is needed to get the optimal rate of convergence $n^{-2\alpha/(2\alpha+1)}$. Indeed, we may apply the results of Section 6.2 to the family of models which satisfies (6.6) and derive an estimator $\tilde{s}$ with a risk bounded by

$$\mathbb{E}_s\left[\|s - \tilde{s}(\boldsymbol{X})\|^2\right] \leq C(\alpha,p) \inf_{J\geq 0} \left[2^{-2J\alpha}\left(|s|_{\alpha,p,\infty}\right)^2 + n^{-1}\|s\|_\infty\left(2^J \vee \log\|s\|_\infty\right)\right].$$

Choosing 2^J of the order of $n^{1/(2\alpha+1)}$ leads to the bound

$$\mathbb{E}_s\left[\|s - \tilde{s}(\boldsymbol{X})\|^2\right] \leq C\left(\alpha, p, |s|_{\alpha,p,\infty}, \|s\|_\infty\right) n^{-2\alpha/(2\alpha+1)},$$

which is valid for all $s \in B^\alpha_{p,\infty}([0,1]) \cap \mathbb{L}_\infty(\mu)$, whatever $\alpha < r$ and p and although α, p, $|s|^\alpha_p$ and $\|s\|_\infty$ are unknown.

6.4.3 Further upper bounds for the risk

When $\alpha \leq 1/p$, i.e. $0 < \delta \leq 1/2$, s may be unbounded and the classical theory does not apply any more. As a consequence the minimax risk over balls in $B^\alpha_{p,\infty}([0,1])$ is presently unknown. Our study will not, unfortunely, solve this problem but, at least, provide some partial information. In this section we assume that $\alpha \leq 1/p$ and, as usual, restrict ourselves to the case $p \leq 2$ so that $\delta \leq \alpha$. We consider the wavelet expansion of s which has been described in Section 6.4.1 and, to avoid unnecessary complications, we also assume that $|s|_{\alpha,p,\infty} \geq 1$. In what follows, the generic constant C (changing from line to line) depends on the choice of the basis and δ.

Since $p \leq 2$,

$$\left(\sum_{k\in\Lambda(j)} \beta_{j,k}^2\right)^{1/2} \leq \left(\sum_{k\in\Lambda(j)} |\beta_{j,k}|^p\right)^{1/p} \leq |s|_{\alpha,p,\infty} 2^{-j(\alpha+1/2-1/p)} = |s|_{\alpha,p,\infty} 2^{-j\delta},$$

hence, for $q \geq -1$ an integer,

$$\left\|\sum_{j>q}\sum_{k\in\Lambda(j)} \beta_{j,k}\varphi_{j,k}\right\| \leq C2^{-q\delta}|s|_{\alpha,p,\infty}. \tag{6.7}$$

The simplest estimators of s are the projection estimators $\hat{s}_q$ over the linear spaces $\overline{S}'_q$ where $\overline{S}'_q$ is spanned by $\{\varphi_{j,k}, -1 \leq j \leq q, k \in \Lambda(j)\}$

$$\hat{s}_q(\boldsymbol{X}) = \sum_{j=-1}^{q}\sum_{k\in\Lambda(j)} \hat{\beta}_{j,k}(\boldsymbol{X})\varphi_{j,k}, \quad \text{with} \quad \hat{\beta}_{j,k}(\boldsymbol{X}) = n^{-1}\sum_{i=1}^{n}\varphi_{j,k}(X_i),$$

The risk of these estimators can be bounded using (2.6) and (6.7) by

$$\mathbb{E}_s\left[\|s - \hat{s}_q(\boldsymbol{X})\|^2\right] \leq d_2^2\left(s, \overline{S}'_q\right) + C2^q/n \leq C'\left[2^{-2q\delta}|s|^2_{\alpha,p,\infty} + 2^q/n\right].$$

A convenient choice of q, depending on δ, then leads to

$$\mathbb{E}_s\left[\|s - \hat{s}_q(\boldsymbol{X})\|^2\right] \leq C|s|^2_{\alpha,p,\infty} n^{-2\delta/(2\delta+1)}.$$

One can actually choose q from the data using a penalized least squares estimator and get a similar risk bound without knowing δ as shown by Theorem 7.5 of Massart (2007). This is the only adaptation result we know for the case $\alpha \leq 1/p$ without the restriction $s \in \overline{\mathbb{L}}_\infty([0,1])$.

Let us now see what our method can do. Since s is a density, it follows from (6.4) and (6.3) that $|\beta_{-1,k}| \leq \|\varphi_{-1,k}\|_\infty \leq K'/\sqrt{2}$, hence

$$\left\|\sum_{k\in\Lambda(-1)} \beta_{-1,k}\varphi_{-1,k}\right\|_\infty \leq \left(K'/\sqrt{2}\right)\left\|\sum_{k\in\Lambda(-1)} \varphi_{-1,k}\right\|_\infty \leq K'^2/2.$$

Moreover, for $j \geq 0$, (6.5) implies that $\sup_{k\in\Lambda(j)} |\beta_{j,k}| \leq 2^{-j\delta}|s|_{\alpha,p,\infty}$. Therefore, by (6.3),

$$\left\|\sum_{k\in\Lambda(j)} \beta_{j,k}\varphi_{j,k}\right\|_\infty \leq C2^{-j(\alpha-1/p)}|s|_{\alpha,p,\infty},$$

and, for $J \geq 0$,

$$\left\|\sum_{j=0}^{J}\sum_{k\in\Lambda(j)} \beta_{j,k}\varphi_{j,k}\right\|_\infty \leq \begin{cases} C|s|_{\alpha,p,\infty} & \text{if } \alpha > 1/p; \\ C(J+1)|s|_{\alpha,p,\infty} & \text{if } \alpha = 1/p; \\ C2^{J(1/p-\alpha)}|s|_{\alpha,p,\infty} & \text{if } \alpha < 1/p. \end{cases}$$

Finally,

$$\left\|\sum_{j=-1}^{J}\sum_{k\in\Lambda(j)}\beta_{j,k}\varphi_{j,k}\right\|_\infty \le C_0L_J|s|_{\alpha,p,\infty} \quad\text{with}\quad L_J=\begin{cases}1 & \text{if } \alpha>1/p;\\ (J+1) & \text{if } \alpha=1/p;\\ 2^{J(1/p-\alpha)} & \text{if } \alpha<1/p.\end{cases}$$

Observing that if $s=u+v$ with $\|u\|_\infty\le z$, then $Q_s(z)\le\|v\|^2$, we can conclude from (6.7) that

$$Q_s\left(C_0L_J|s|_{\alpha,p,\infty}\right)\le C'2^{-2J\delta}|s|^2_{\alpha,p,\infty}.$$

Let us now turn back to the family of linear models described in Section 6.4.1 that satisfy (6.6). Theorem 6 asserts the existence of an estimator $\tilde{s}(\boldsymbol{X})$ based on this family of models and satisfying

$$\mathbb{E}_s\left[\|s-\tilde{s}(\boldsymbol{X})\|^2\right]\le C\inf_{z\ge2}\inf_{m\in\mathcal{M}}\left[d_2^2\left(s,\overline{S}_m\right)+\frac{z\left(\overline{D}_m\vee\Delta_m\vee\log z\right)}{n}+Q_s(z)\right].$$

Given the integers J,J', we may set $z=z_{J'}=C_0L_{J'}|s|_{\alpha,p,\infty}$ and restrict the minimization to $m\in\mathcal{M}_J$ which leads to

$$\mathbb{E}_s\left[\|s-\tilde{s}(\boldsymbol{X})\|^2\right]\le C\left[|s|^2_{\alpha,p,\infty}\left(2^{-2J\alpha}+2^{-2J'\delta}\right)+n^{-1}L_{J'}|s|_{\alpha,p,\infty}\left(2^J\vee\log z_{J'}\right)\right].$$

Since $L_{J'}\left(2^J\vee\log z_{J'}\right)$ is a nondecreasing function of both J and J', this last bound is optimized when $J\alpha$ and $J'\delta$ are approximately equal which leads to choosing the integer J' so that $J\alpha/\delta\le J'<J\alpha/\delta+1$, hence $2^{-2J'\delta}\le2^{-2J\alpha}$. Assuming, moreover, that $2^J\ge\log|s|_{\alpha,p,\infty}$, which implies that $2^J\ge C'\log z_{J'}$, we get

$$\mathbb{E}_s\left[\|s-\tilde{s}(\boldsymbol{X})\|^2\right]\le C|s|^2_{\alpha,p,\infty}\left[2^{-2J\alpha}+2^J\left(n|s|_{\alpha,p,\infty}\right)^{-1}L_{J'}\right].$$

We finally fix J so that $2^J\ge G>2^{J-1}$, where G is defined below. This choice ensures that $G\ge\log|s|_{\alpha,p,\infty}$ for n large enough (depending on $|s|_{\alpha,p,\infty}$), which we assume here.

— If $\alpha>1/p$ we set $G=\left(n|s|_{\alpha,p,\infty}\right)^{1/(2\alpha+1)}$ which leads to a risk bound of the form

$$Cn^{-2\alpha/(2\alpha+1)}\left(|s|_{\alpha,p,\infty}\right)^{(2\alpha+2)/(2\alpha+1)}.$$

— If $\alpha=1/p$, $L'_J<J\alpha/\delta+2$ and we take $G=\left(n|s|_{\alpha,p,\infty}/\log n\right)^{1/(2\alpha+1)}$ which leads to the risk bound

$$C(n/\log n)^{-2\alpha/(2\alpha+1)}\left(|s|_{\alpha,p,\infty}\right)^{(2\alpha+2)/(2\alpha+1)}.$$

— Finally, for $\alpha<1/p$, $L_{J'}<\sqrt{2}\,2^{(J\alpha/\delta)(1/p-\alpha)}$ and we set $G=\left(n|s|_{\alpha,p,\infty}\right)^{1/[\alpha+1+\alpha/(2\delta)]}$ which leads to the bound

$$Cn^{-2\alpha/[\alpha+1+\alpha/(2\delta)]}\left(|s|_{\alpha,p,\infty}\right)^{(2+(\alpha/\delta)/[\alpha+1+\alpha/(2\delta)]}.$$

6.4.4 Some lower bounds

Lower bounds of the form $n^{-2\alpha/(1+2\alpha)}$ for the minimax risk on Besov balls are well-known (deriving from lower bounds for Hölder spaces) and they are sharp for $\alpha > 1/p$, as shown in Donoho, Johnstone, Kerkyacharian and Picard (1996). To derive new lower bounds for the case $\alpha < 1/p$ we introduce some probability density $f \in B^{\alpha}_{p,\infty}([0,1])$ with compact support included in $(0,1)$ and Besov semi-norm $|f|^{\alpha}_{p}$. Then we set $g(x) = af(2anx)$ for some $a > (2n)^{-1}$ to be fixed later. Then $g(x) = 0$ for $x \notin \left(0,(2an)^{-1}\right)$,

$$\|g\|_q = a(2an)^{-1/q}\|f\|_q \qquad \text{and} \qquad |g|^{\alpha}_{p} = a(2an)^{\alpha-1/p}|f|^{\alpha}_{p}.$$

Let us now set $t = g + \left[1-(2n)^{-1}\right]\mathbb{1}_{[0,1]}$, so that t is a density belonging to $B^{\alpha}_{p,\infty}([0,1])$ with Besov semi-norm

$$|t|^{\alpha}_{p} = |g|^{\alpha}_{p} = Ka^{1+\alpha-1/p}n^{\alpha-1/p} \quad \text{with } K = 2^{\alpha-1/p}|f|^{\alpha}_{p}.$$

For a given value of the constant $K' > 0$, the choice $a = \left[K'n^{1/p-\alpha}\right]^{1/(1+\alpha-1/p)} > (2n)^{-1}$ (at least for n large) leads to $|t|^{\alpha}_{p} = KK'$ so that K' determines $|t|^{\alpha}_{p}$. We also consider the density $u(x) = t(1-x)$ which has the same Besov semi-norm. Then

$$h^2(t,u) = \int_0^{(2an)^{-1}} \left(\sqrt{g+[1-(2n)^{-1}]} - \sqrt{1-(2n)^{-1}}\right)^2 < \int_0^{(2an)^{-1}} g = (2n)^{-1},$$

and it follows from Le Cam (1973) that any estimator $\hat{s}$ based on n i.i.d. observations satisfies

$$\max\left\{\mathbb{E}_t\left[\|t-\hat{s}\|^2\right], \mathbb{E}_u\left[\|u-\hat{s}\|^2\right]\right\} \geq C\|t-u\|^2 = 2C\|g\|^2 = Can^{-1}\|f\|^2.$$

Since $an^{-1} = K'^{1/(\delta+1/2)}n^{-2\delta/(\delta+1/2)}$, we finally get

$$\max\left\{\mathbb{E}_t\left[\|t-\hat{s}\|^2\right], \mathbb{E}_u\left[\|u-\hat{s}\|^2\right]\right\} \geq C\left(|t|^{\alpha}_{p}\right)^{2/(2\delta+1)} n^{-4\delta/(2\delta+1)},$$

where C depends on K', $\|f\|$, $|f|^{\alpha}_{p}$ and δ. One can check that this rate is slower than $n^{-2\alpha/(1+2\alpha)}$, when $0 < \delta < \alpha[2(\alpha+1)]^{-1}$ or, equivalently, when $\alpha + [2(\alpha+1)]^{-1} < 1/p$.

6.4.5 Conclusion

In the case $\alpha > 1/p$, the estimator that we built in Section 6.4.3 has the usual rate of convergence with respect to n, namely $n^{-2\alpha/(2\alpha+1)}$, which is known to be optimal, and we can extend the result to the borderline case $\alpha = 1/p$ with only a logarithmic loss. The situation is different when $\alpha < 1/p$ for which, to our knowledge, the value of the minimax risk is still unknown. The rate $n^{-2\alpha/[\alpha+1+\alpha/(2\delta)]}$ that we get is worse than the one valid for $\alpha > 1/p$ and also than the lower bound $n^{-4\delta/(2\delta+1)}$ that we derived in the previous section. It can be compared with the risk of the penalized least squares estimators based on the nested models $\overline{S}'_q$, which is, as we have seen, bounded by $Cn^{-2\delta/(2\delta+1)}$. Our rate is better when $\alpha > 2\delta/(2\delta+1)$, which is always true for $\alpha \geq 1/2$ since $\delta < 1/2$. When $\alpha < 1/2$, hence $p > 2/(2\alpha+1) > 1$, it also holds for $p < 2(1-\alpha)/\left(1-2\alpha^2\right) < \alpha^{-1}$. We are convinced that our rate is always suboptimal in the range $\alpha \leq 1/p$ but are presently unable to derive the correct minimax rate.

7 Proofs

7.1 Proof of Proposition 2

For simplicity, we shall write $h(\theta,\lambda)$ for $h(s_\theta,s_\lambda)$ and analogously $d_2(\theta,\lambda)$ for $d_2(s_\theta,s_\lambda)$. Let us first evaluate $h^2(\theta,\lambda)$ for $0<\theta<\lambda\le 1/3$. Setting $\beta_\theta=\left(\theta^2+\theta+1\right)^{-1}\in[9/13,1)$, we get

$$\begin{aligned}
2h^2(\theta,\lambda) &= \int_0^1\left(\sqrt{s_\theta(x)}-\sqrt{s_\lambda(x)}\right)^2dx\\
&= \theta^3\left(\theta^{-1}-\lambda^{-1}\right)^2+\left(\lambda^3-\theta^3\right)\left(\lambda^{-1}-\sqrt{\beta_\theta}\right)^2+\left(1-\lambda^3\right)\left(\sqrt{\beta_\theta}-\sqrt{\beta_\lambda}\right)^2\\
&= (\lambda-\theta)\frac{\theta}{\lambda}\left(1-\frac{\theta}{\lambda}\right)+(\lambda-\theta)\left[1+\frac{\theta}{\lambda}+\left(\frac{\theta}{\lambda}\right)^2\right]\left(1-\lambda\sqrt{\beta_\theta}\right)^2\\
&\quad+\left(1-\lambda^3\right)\left(\sqrt{\beta_\theta}-\sqrt{\beta_\lambda}\right)^2.
\end{aligned}$$

Note that the monotonicity of $\theta\mapsto\beta_\theta$ implies that

$$4/9<\left(1-\lambda\sqrt{\beta_\theta}\right)^2<1,\qquad \sqrt{\beta_\theta}+\sqrt{\beta_\lambda}>2\sqrt{\beta_{1/3}}=6/\sqrt{13}$$

and

$$0<\beta_\theta-\beta_\lambda=\frac{(\lambda-\theta)(\lambda+\theta+1)}{(\theta^2+\theta+1)(\lambda^2+\lambda+1)}<\lambda-\theta. \tag{7.1}$$

It follows that

$$0<\left(\sqrt{\beta_\theta}-\sqrt{\beta_\lambda}\right)^2=\frac{(\beta_\theta-\beta_\lambda)^2}{\left(\sqrt{\beta_\theta}+\sqrt{\beta_\lambda}\right)^2}<\frac{13}{36}(\lambda-\theta)^2=\frac{13\lambda}{36}(\lambda-\theta)\left(1-\frac{\theta}{\lambda}\right)$$

and

$$0<\left(1-\lambda^3\right)\left(\sqrt{\beta_\theta}-\sqrt{\beta_\lambda}\right)^2<\frac{13\lambda\left(1-\lambda^3\right)}{36}(\lambda-\theta)\left(1-\frac{\theta}{\lambda}\right)<\frac{2(\lambda-\theta)}{17}\left(1-\frac{\theta}{\lambda}\right).$$

We can therefore write

$$G=2(\lambda-\theta)^{-1}h^2(\theta,\lambda)=z(1-z)+c_1(\theta,\lambda)\left(1+z+z^2\right)+c_2(\theta,\lambda)(1-z),$$

with $z=\theta/\lambda\in(0,1)$, $4/9<c_1(\theta,\lambda)<1$ and $0<c_2(\theta,\lambda)<2/17$. Since, for given values of c_1 and c_2, the right-hand side is increasing with respect to z, $4/9<c_1<G<3c_1<3$ and we conclude that for all θ and λ in $(0,1/3]$,

$$h^2(\theta,\lambda)=C(\theta,\lambda)|\theta-\lambda|\quad\text{with } 2/9<C(\theta,\lambda)<3/2.$$

It immediately follows that the set $S_\eta=\{s_{\lambda_j},\ j\ge 0\}$ with $\lambda_j=(2j+1)2\eta^2/3$ is an η-net for the family $\overline{S}$. On the other hand, given $\lambda\in(0,1/3)$ and $r\ge 2\eta$, in order that $s_{\lambda_j}\in\mathcal{B}(s_\lambda,r)$, it is required that $h^2(\lambda_j,\lambda)=C(\lambda_j,\lambda)|\lambda_j-\lambda|<r^2$ which implies that $|\lambda_j-\lambda|<(9/2)r^2$ and therefore

$$|S_\eta\cap\mathcal{B}(s_\lambda,r)|\le 1+(27/4)(r/\eta)^2\le\exp\left[0.84(r/\eta)^2\right]\quad\text{for all } s_\lambda\in\overline{S}.$$

It follows from Lemma 2 of Birgé (2006a) that $\overline{S}$ has a metric dimension bounded by 3.4 and Corollary 3 of Birgé (2006a) implies that a suitable T-estimator $\tilde{s}$ built on S_η has a risk satisfying

$$\mathbb{E}_{s_\theta}\left[h^2(s,\tilde{s})\right] \leq Cn^{-1} \quad \text{for all } s_\theta \in \overline{S}.$$

Let us now proceed with the $\mathbb{L}_2$-distance d_2.

$$\begin{aligned}
d_2^2(\theta,\lambda) &= \theta^3\left(\theta^{-2}-\lambda^{-2}\right)^2 + \left(\lambda^3-\theta^3\right)\left(\lambda^{-2}-\beta_\theta\right)^2 + \left(1-\lambda^3\right)\left(\beta_\theta-\beta_\lambda\right)^2 \\
&= \left(\frac{1}{\theta}-\frac{1}{\lambda}\right)\left(1-\frac{\theta}{\lambda}\right)\left(1+\frac{\theta}{\lambda}\right)^2 \\
&\quad + \left(\frac{1}{\theta}-\frac{1}{\lambda}\right)\left[\frac{\theta}{\lambda}+\left(\frac{\theta}{\lambda}\right)^2+\left(\frac{\theta}{\lambda}\right)^3\right]\left(1-\lambda^2\beta_\theta\right)^2 \\
&\quad + \left(\frac{1}{\theta}-\frac{1}{\lambda}\right)\left(1-\frac{\theta}{\lambda}\right)\theta\lambda^2\left(1-\lambda^3\right)\left(\frac{\beta_\theta-\beta_\lambda}{\lambda-\theta}\right)^2.
\end{aligned}$$

Since $8/9 < 1-\lambda^2\beta_\theta < 1$ and, by (7.1),

$$0 < \theta\lambda^2\left(1-\lambda^3\right)\left(\frac{\beta_\theta-\beta_\lambda}{\lambda-\theta}\right)^2 < \frac{1}{27},$$

we conclude that

$$G = \left(\theta^{-1}-\lambda^{-1}\right)^{-1}d_2^2(\theta,\lambda) = (1-z)(1+z)^2 + c_1(\theta,\lambda)\left(z+z^2+z^3\right) + c_2(\theta,\lambda)(1-z),$$

with $z=\theta/\lambda \in (0,1)$, $8/9 < c_1(\theta,\lambda) < 1$ and $0 < c_2(\theta,\lambda) < 1/27$. It follows that

$$1 < 1+z-z^2-z^3+(8/9)\left(z+z^2+z^3\right) < G < 1+2z+(1/27)(1-z) < 3,$$

which finally implies that, for all θ and λ in $(0,1/3]$,

$$d_2^2(\theta,\lambda) = C(\theta,\lambda)\left|\theta^{-1}-\lambda^{-1}\right| \quad \text{with } 1 < C(\theta,\lambda) < 3.$$

Now setting $S_\eta = \{s_{\lambda_j},\ j \geq 0\}$ with $\lambda_j = \left(3+2j\eta^2/3\right)^{-1}$ we deduce as before that S_η is an η-net for $\overline{S}$. In order that $s_{\lambda_j} \in \mathcal{B}(s_\lambda, x\eta)$, it is required that $d_2^2(\lambda_j,\lambda) = C(\theta,\lambda)|\lambda_j^{-1}-\lambda^{-1}| < x^2\eta^2$, which implies that $|\lambda_j^{-1}-\lambda^{-1}| < x^2\eta^2$. It follows that the number of elements of S_η contained in the ball is bounded by $3x^2/2+1 \leq \exp\left(x^2/2\right)$ for $x \geq 2$. Hence the metric dimension of $\overline{S}$ with respect to the $\mathbb{L}_2$-distance is bounded by 2. It nevertheless follows from the fact that $h(\theta,\lambda) \to 0$ while $d_2(\theta,\lambda) \to +\infty$ when θ and λ tend to zero and classical arguments of Le Cam (1973) — see also Donoho and Liu (1987) or Yu (1997) — that the minimax risk over $\overline{S}$ is infinite when we use the $\mathbb{L}_2$-loss.

7.2 Proof of Lemma 1

Let us begin with a preliminary lemma.

Lemma 2 *Let F and G be two disjoint sets with positive measures $\alpha = \mu(F)$ and $\beta = \mu(G)$ and $g \in \overline{\mathbb{L}}_2$ such that $\inf_{x\in F} g(x) > 0$. Set $g_\varepsilon = g + \varepsilon(\alpha 1\!\!1_G - \beta 1\!\!1_F)$ for $\varepsilon > 0$. Then g_ε is a density for ε small enough and for any $f \in \overline{\mathbb{L}}_2$,*

$$\lim_{\varepsilon\to 0} \frac{1}{2\varepsilon}\left[d_2^2(g_\varepsilon, f) - d_2^2(g, f)\right] = \alpha \int_G (g-f)\,d\mu - \beta \int_F (g-f)\,d\mu \tag{7.2}$$

and

$$\lim_{\varepsilon\to 0} \frac{2}{\varepsilon}\left[h^2(g_\varepsilon, f) - h^2(g, f)\right] = \beta \int_F \sqrt{fg^{-1}}\,d\mu - \alpha \int_G \sqrt{fg^{-1}}\,d\mu, \tag{7.3}$$

with the convention that $\int_G \sqrt{fg^{-1}}\,d\lambda = +\infty$ if either $\mu(G \cap \{g = 0\} \cap \{f > 0\}) > 0$ or the integral diverges.

Proof: Since $\int g_\varepsilon\,d\mu = 1$ and $g_\varepsilon \geq 0$ for ε small enough g_ε is a density. Moreover, setting $k = \alpha 1\!\!1_G - \beta 1\!\!1_F$, we get

$$d_2^2(g_\varepsilon, f) = \int (g + \varepsilon k - f)^2\,d\mu = d_2^2(g, f) + \varepsilon^2 \|k\|^2 + 2\varepsilon \int k(g-f)\,d\mu$$

and (7.2) follows. Let $\Delta(\varepsilon) = \varepsilon^{-1}\left[h^2(g_\varepsilon, f) - h^2(g, f)\right]$ and fix $\eta > 0$. Then

$$\begin{aligned}
\Delta(\varepsilon) &= \varepsilon^{-1}\left[\int \sqrt{gf}\,d\mu - \int \sqrt{(g+\varepsilon k)f}\,d\mu\right] \\
&= \varepsilon^{-1}\left[\int_F \left[\sqrt{gf} - \sqrt{(g-\varepsilon\beta)f}\right] d\mu + \int_G \left[\sqrt{gf} - \sqrt{(g+\varepsilon\alpha)f}\right] d\mu\right] \\
&= \int_F \frac{\beta\sqrt{f}}{\sqrt{g-\varepsilon\beta} + \sqrt{g}}\,d\mu - \int_{G\cap\{g>0\}} \frac{\alpha\sqrt{f}}{\sqrt{g+\varepsilon\alpha} + \sqrt{g}}\,d\mu \\
&\quad - \int_{G\cap\{g=0\}\cap\{f>0\}} \sqrt{\alpha f/\varepsilon}\,d\mu.
\end{aligned}$$

When ε tends to 0, the first integral converges to $(\beta/2)\int_F \sqrt{fg^{-1}}\,d\mu$ and the second one converges to $(\alpha/2)\int_{G\cap\{g>0\}} \sqrt{fg^{-1}}\,d\mu$, by monotone convergence. The last one converges to $+\infty$ if $\mu(G \cap \{g = 0\} \cap \{f > 0\}) > 0$ and 0 otherwise, which achieves the proof of (7.3). □

If $\|v_1 \vee v_2\|_\infty > 2B$, we may assume, exchanging the roles of v_1 and v_2 if necessary, that $\mu(A) > 0$ with $A = \{v_1 \geq v_2 \text{ and } v_1 > 2B\}$. Let $C = \{v_1 < B \wedge v_2\}$. If $\mu(C) > 0$, we may apply Lemma 2 with $F = A$, $G = C$, $g = v_1$ and $v_1' = g_\varepsilon$. We first set $f = t$. Since $v_1 - t < B$ on C while $v_1 - t > B$ on A, it follows from (7.2) that $d_2(v_1', t) < d_2(v_1, t)$ for ε small enough. If we now set $f = v_2$ and use (7.3), we see that $h(v_1', v_2) < h(v_1, v_2)$ since $v_2 \leq v_1$ on A and $v_2 > v_1$ on C. We conclude by setting $v_2' = v_2$. If $\mu(C) = 0$, then $\mu(\{B \leq v_1 < v_2\}) + \mu(\{v_2 \leq v_1 < B\}) = 1$ and both sets have positive μ-measure since $v_1 \neq v_2$. In this case we set $F = \{B \leq v_1 < v_2\}$, $G = \{v_2 \leq v_1 \wedge u\}$ and $g = v_2$. Then $\mu(F) > 0$ and $\mu(G) > 0$ since $u \leq B < v_2$ on F and they are densities. If we use (7.2) with $f = u$, we derive that $d_2(v_2', u) < d_2(v_2, u)$ for ε small enough and if we use (7.3) with $f = v_1$, we derive that $h(v_2', v_1) < h(v_2, v_1)$, in which case we set $v_1' = v_1$.

7.3 Proof of Theorem 5

We consider the family of tests $\psi(t_n,t_p,\boldsymbol{X})=\psi_{t_n,t_p,x}(\boldsymbol{X})$ provided by the assumption with $x=A|p-n|$. Given this family of tests and $S=\{t_i,i\ge 1\}$, we define the random function $\mathcal{D}_{\boldsymbol{X}}$ on S as in Birgé (2006a), i.e. we set $\mathcal{R}_i=\{t_j\in S, j\neq i\,|\,\psi(t_i,t_j,\boldsymbol{X})=t_j\}$ and

$$\mathcal{D}_{\boldsymbol{X}}(t_i)=\begin{cases}\sup\limits_{t_j\in\mathcal{R}_i}\left\{d(t_i,t_j)\right\} & \text{if } \mathcal{R}_i\neq\emptyset;\\ 0 & \text{if } \mathcal{R}_i=\emptyset.\end{cases}\tag{7.4}$$

Given some $t_i\in S$, we want to bound

$$\mathbb{P}_s\left[\mathcal{D}_{\boldsymbol{X}}(t_i)>xy_i\right]\quad\text{ for } x\ge 1\quad\text{ and }\quad y_i=4d(s,t_i)\vee\sqrt{Aa^{-1}i2^i}.$$

Let us define the integer K by $x^2<2^K\le 2x^2$. Then

$$K\ge 1,\quad a2^{-i-K}(xy_i)^2\ge a2^{-i-1}y_i^2\ge Ai/2\quad\text{and}\quad e^{-AK}\le x^{-2A/\log 2}.\tag{7.5}$$

Now, setting $y=xy_i$, observe that

$$\mathbb{P}_s\left[\mathcal{D}_{\boldsymbol{X}}(t_i)>y\right]=\mathbb{P}_s\left[\,\exists j\text{ with } d(t_i,t_j)>y\text{ and }\psi(t_i,t_j,\boldsymbol{X})=t_j\right]\le\Sigma_1+\Sigma_2,$$

with

$$\Sigma_1=\sum_{j<i}\mathbb{1}_{d(t_i,t_j)>y}\,\mathbb{P}_s\left[\psi(t_i,t_j,\boldsymbol{X})=t_j\right];\quad \Sigma_2=\sum_{j>i}\mathbb{1}_{d(t_i,t_j)>y}\,\mathbb{P}_s\left[\psi(t_i,t_j,\boldsymbol{X})=t_j\right].$$

If $i=1$, then $\Sigma_1=0$ and if $i\ge 2$, we can use (5.2) and $y\ge 4d(s,t_i)$ to derive that

$$\begin{aligned}\Sigma_1 &\le B\sum_{j<i}\mathbb{1}_{d(t_i,t_j)>y}\,\exp\left[-a2^{-i}d^2(t_i,t_j)+A|i-j|\right]\\ &\le B\exp\left[-a2^{-i}y_i^2x^2+Ai\right]\sum_{j\ge 1}e^{-Aj}\\ &\le B\frac{e^{-A}}{1-e^{-A}}\exp\left[-Ai\left(x^2-1\right)\right]\;\le\; B\frac{e^{-A}}{1-e^{-A}}\exp\left[-A\left(x^2-1\right)\right]\\ &\le B\left(1-e^{-A}\right)^{-1}\exp\left[-Ax^2\right]\;\le\; B\left(1-e^{-A}\right)^{-1}x^{-2A/\log 2},\end{aligned}$$

where we used (7.5), $i\ge 1$ and $x\ge 1$. Also, by (5.1),

$$\begin{aligned}\Sigma_2 &\le B\sum_{j>i}\mathbb{1}_{d(t_i,t_j)>y}\,\exp\left[-a2^{-j}d^2(t_i,t_j)-A|i-j|\right]\\ &\le B\sum_{j>i}\exp\left[-a2^{-j}y^2-A(j-i)\right]\;=\;B\sum_{k=1}^{+\infty}\exp\left[-a2^{-i-k}y^2-Ak\right]\\ &\le B\left[\sum_{k=1}^{K}\exp\left[-a2^{-i-k}y^2-Ak\right]+\sum_{k>K}\exp[-Ak]\right]\;=\;B(\Sigma_3+\Sigma_4),\end{aligned}$$

with $\Sigma_4=e^{-AK}\left(e^A-1\right)^{-1}$ and, by (7.5),

$$\begin{aligned}\Sigma_3 &= e^{-AK}\sum_{j=0}^{K-1}\exp\left[-a2^{-i-K+j}y^2+Aj\right]\;\le\; e^{-AK}\sum_{j=0}^{K-1}\exp\left[-A(i2^{j-1}-j)\right]\\ &\le e^{-AK}\sum_{j\ge 0}\exp\left[-\left(2^{j-1}-j\right)\right]\;<\;3e^{-AK}.\end{aligned}$$

We finally get, putting all the bounds together and using (7.5) again,

$$\mathbb{P}_s\left[\mathcal{D}_{\boldsymbol{X}}(t_i) > xy_i\right] \leq BC(A)x^{-2A/\log 2} \quad \text{for } x \geq 1. \tag{7.6}$$

As a consequence $\mathcal{D}_{\boldsymbol{X}}(t_i) < +\infty$ a.s. and we can define

$$\hat{s}_A = t_p \quad \text{with } p = \min\left\{j \;\middle|\; \mathcal{D}_{\boldsymbol{X}}(t_j) < \inf_i \mathcal{D}_{\boldsymbol{X}}(t_i) + \sqrt{Aa^{-1}}\right\}.$$

In view of the definition of $\mathcal{D}_{\boldsymbol{X}}$, $d(t_i,t_j) \leq \mathcal{D}_{\boldsymbol{X}}(t_i) \vee \mathcal{D}_{\boldsymbol{X}}(t_j)$, hence, for all $t_i \in S$, $d\left(\hat{s}_A, t_i\right) \leq \mathcal{D}_{\boldsymbol{X}}(t_i) + \sqrt{Aa^{-1}}$ and

$$d\left(\hat{s}_A, s\right) \leq \mathcal{D}_{\boldsymbol{X}}(t_i) + \sqrt{Aa^{-1}} + d(s,t_i) < \mathcal{D}_{\boldsymbol{X}}(t_i) + y_i.$$

It then follows from (7.6) that

$$\mathbb{P}_s\left[d\left(\hat{s}_A, s\right) > zy_i\right] \leq BC(A)(z-1)^{-2A/\log 2} \quad \text{for } z \geq 2.$$

Integrating with respect to z leads to

$$\mathbb{E}_s\left[\left(d\left(\hat{s}_A, s\right)/y_i\right)^q\right] \leq BC(A,q) \quad \text{for } 1 \leq q < 2A/\log 2,$$

and, since t_i is arbitrary in S,

$$\mathbb{E}_s\left[d^q\left(\hat{s}_A, s\right)\right] \leq BC(A,q)\inf_{i\geq 1}\left[d^q(s,t_i) \vee \left(a^{-1}i2^i\right)^{q/2}\right] \quad \text{for } 1 \leq q < 2A/\log 2.$$

References

BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrscheinlichkeitstheorie Verw. Geb.* **65**, 181-237.

BIRGÉ, L. (2004). Model selection for Gaussian regression with random design. *Bernoulli* **10**, 1039 -1051.

BIRGÉ, L. (2006a). Model selection via testing : an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. Henri Poincaré Probab. et Statist.* **42**, 273-325.

BIRGÉ, L. (2006b). Statistical estimation with model selection. *Indagationes Mathematicae* **17**, 497-537.

BIRGÉ, L. (2007). Model selection for Poisson processes, in *Asymptotics: particles, processes and inverse problems, Festschrift for Piet Groeneboom* (E. Cator, G. Jongbloed, C. Kraaikamp, R. Lopuhaä and J. Wellner, eds). IMS Lecture Notes – Monograph Series **55**, 32-64.

BIRGÉ, L. and MASSART, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics* (D. Pollard, E. Torgersen and G. Yang, eds.), 55-87. Springer-Verlag, New York.

BIRGÉ, L. and MASSART, P. (2000). An adaptive compression algorithm in Besov spaces. *Constructive Approximation* **16** 1-36.

BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* **3**, 203-268.

BIRGÉ, L. and ROZENHOLC, Y. (2006). How many bins should be put in a regular histogram. *ESAIM-Probab. & Statist.* **10**, 24-45.

CENCOV, N.N. (1962). Evaluation of an unknown distribution density from observations. *Soviet Math.* **3**, 1559-1562.

DELYON, B. and JUDITSKY, A. (1996). On minimax wavelet estimators. *Appl. Comput. Harmonic Anal.* **3**, 215-228.

DeVORE, R.A. and LORENTZ, G.G. (1993). *Constructive Approximation.* Springer-Verlag, Berlin.

DEVROYE, L. (1987). *A Course in Density Estimation.* Birkhäuser, Boston.

DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric Density Estimation: The L_1 View.* John Wiley, New York.

DONOHO, D.L., JOHNSTONE, I.M., KERKYACHARIAN, G. and PICARD, D. (1996). Density estimation by wavelet thresholding *Ann. Statist.* **24**, 508-539.

DONOHO, D.L. and LIU, R.C. (1987). Geometrizing rates of convergence I. Technical report 137. Department of Statistics, University of California, Berkeley.

JUDITSKY, A., RIGOLLET, P. and TSYBAKOV, A. (2007). Learning by mirror averaging. To appear in *Ann. Statist.*

KERKYACHARIAN, G. and PICARD, D. (2000). Thresholding algorithms, maxisets and well-concentrated bases. *Test* **9**, 283-344.

Le CAM, L.M. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** , 38-53.

Le CAM, L.M. (1975). On local and global properties in the theory of asymptotic normality of experiments. *Stochastic Processes and Related Topics, Vol. 1* (M. Puri, ed.), 13-54. Academic Press, New York.

Le CAM, L.M. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, New York.

LOUNICI, K. (2008). Aggregation of density estimators for the L^π risk with $1 \leq \pi < \infty$. Unpublished manuscript.

MASSART, P. (2007). Concentration Inequalities and Model Selection. In *Lecture on Probability Theory and Statistics, Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003* (J. Picard, ed.). Lecture Note in Mathematics, Springer-Verlag, Berlin.

RIGOLLET, T. (2006). Ph.D. thesis, University Pierre et Marie Curie, Paris.

RIGOLLET, T. and TSYBAKOV, A.B. (2007). Linear and convex aggregation of density estimators. *Math. Methods of Statis.* **16**, 260-280.

SAMAROV, A. and TSYBAKOV, A.B. (2007) Aggregation of density estimators and dimension reduction. *Advances in Statistical Modeling and Inference. Essays in Honor of Kjell A. Doksum* (V.Nair, ed.), World Scientific, Singapore e.a., 233-251.

YANG, Y. (2000). Mixing strategies for density estimation. *Ann. Statist.* **28**, 75-87.

YANG, Y. and BARRON, A.R. (1998). An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory* **44**, 95-116.

YU, B. (1997). Assouad, Fano and Le Cam. In *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics* (D. Pollard, E. Torgersen and G. Yang, eds.), 423-435. Springer-Verlag, New York.

Lucien BIRGÉ
UMR 7599 "Probabilités et modèles aléatoires"
Laboratoire de Probabilités, boîte 188
Université Paris VI, 4 Place Jussieu
F-75252 Paris Cedex 05

France

e-mail: lucien.birge@upmc.fr